ELEMENTARY TREATISE

ON

NATURAL PHILOSOPHY.

BY

A. PRIVAT-DESCHANEL,

FORMERLY PROFESSOR OF PHYSICS IN THE LYCÉE LOUIS-LE-GRAND,
INSPECTOR OF THE ACADEMY OF PARIS.

TRANSLATED AND EDITED, WITH EXTENSIVE MODIFICATIONS,

BY J. D. EVERETT, M. A., D. C. L., F. R. S., F. R. S. E.,

PROFESSOR OF NATURAL PHILOSOPHY IN THE QUEEN'S COLLEGE, BELFAST.

IN FOUR PARTS.

PART IV.

SOUND AND LIGHT.

ILLUSTRATED BY

192 ENGRAVINGS ON WOOD, AND ONE COLORED PLATE.

REVISED EDITION.

NEW YORK:

D. APPLETON AND COMPANY,

1, 3, AND 5 BOND STREET.

1884.

PREFACE TO THE PRESENT EDITION OF PART IV.

THE present edition of this Part contains no radical changes. The numbering of the chapters and sections has been altered to make it consecutive with the other three Parts, but there has been no re-arrangement.

Additions have been made under the following heads (those marked with an asterisk were introduced in a previous edition):—

Mathematical note on stationary undulation;
Edison's phonograph;
Michelson's measurement of the velocity of light;
Astronomical refraction;
*Refraction at a spherical surface;
Refraction through a sphere;
Brightness of image on screen;
Field of view in telescope;
*Curved rays of sound;
*Retardation-gratings and reflection-gratings;
Kerr's magneto-optic discoveries;

besides briefer additions and emendations which it would be tedious to enumerate. The whole volume has been minutely revised; and a copious collection of examples arranged in order, with answers, has been substituted for the original collection of "Problems."

J. D. E.

BELFAST, *May*, 1881.

NOTE PREFIXED TO FIRST EDITION.

IN the present Part, the chapters relating to Consonance and Dissonance, Colour, the Undulatory Theory, and Polarization, are

the work of the Editor; besides numerous changes and additions in other places.

The numbering of the original sections has been preserved only to the end of Chapter LX.; the two last chapters of the original having been transposed for greater convenience of treatment. With this exception, the announcements made in the "Translator's Preface," at the beginning of Part I., are applicable to the entire work.

CONTENTS—PART IV.

THE NUMBERS REFER TO THE SECTIONS.

ACOUSTICS.

CHAPTER LXII. PRODUCTION AND PROPAGATION OF SOUND.

CHAPTER LXIII. NUMERICAL EVALUATION OF SOUND.

CHAPTER LXIV. MODES OF VIBRATION.

OPTICS.

CHAPTER LXX. LENSES.

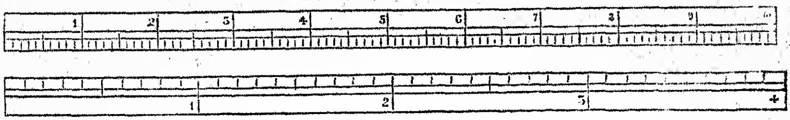
CHAPTER LXXI. VISION AND OPTICAL INSTRUMENTS.

CHAPTER LXXII. DISPERSION. STUDY OF SPECTRA.

CHAPTER LXXIII. COLOUR.

FRENCH AND ENGLISH MEASURES.

A DECIMETRE DIVIDED INTO CENTIMETRES AND MILLIMETRES.



INCHES AND TENTHS.

REDUCTION OF FRENCH TO ENGLISH MEASURES.

Length.

1 millimetre $=\cdot03937$ inch, or about $\frac{1}{25}$ inch.
1 centimetre $=\cdot3937$ inch.
1 decimetre $=3\cdot937$ inch.
1 metre $=39\cdot37$ inch $=3\cdot281$ ft. $=1\cdot0936$ yd.
1 kilometre $=1093\cdot6$ yds., or about $\frac{5}{8}$ mile.
More accurately, 1 metre $=39\cdot370432$ in. $=3\cdot2808693$ ft. $=1\cdot09362311$ yd.

Area.

1 sq. millim. $=\cdot00155$ sq. in.
1 sq. centim. $=\cdot155$ sq. in.
1 sq. decim. $=15\cdot5$ sq. in.
1 sq. metre $=1550$ sq. in. $=10\cdot764$ sq. ft. $=1\cdot196$ sq. yd.

Volume.

1 cub. millim. $=\cdot000061$ cub. in.
1 cub. centim. $=\cdot061025$ cub. in.
1 cub. decim. $=61\cdot0254$ cub. in.
cub. metre $=61025$ cub. in. $=35\cdot3156$ cub. ft. $=1\cdot308$ cub. yd.

The Litre (used for liquids) is the same as the cubic decimetre, and is equal to $1\cdot7617$ pint, or $\cdot22021$ gallon.

Mass and Weight.

1 milligramme $=\cdot01543$ grain.
1 gramme $=15\cdot432$ grain.
1 kilogramme $=15432$ grains $=2\cdot205$ lbs. avoir.
More accurately, the kilogramme is $2\cdot20462125$ lbs.

Miscellaneous.

1 gramme per sq. centim. $=2\cdot0481$ lbs. per sq. ft.
1 kilogramme per sq. centim. $=14\cdot223$ lbs. per sq. in.
1 kilogrammetre $=7\cdot2331$ foot-pounds.
1 force de cheval $=75$ kilogrammetres per second, or $542\frac{1}{2}$ foot-pounds per second nearly, whereas 1 horse-power (English) $=550$ foot-pounds per second.

REDUCTION TO C.G.S. MEASURES. (See page 48.)

[*cm.* denotes centimetre(s); *gm.* denotes gramme(s).]

Length.

1 inch $=2\cdot54$ centimetres, nearly.
1 foot $=30\cdot48$ centimetres, nearly.
1 yard $=91\cdot44$ centimetres, nearly.
1 statute mile $=160933$ centimetres, nearly.
More accurately, 1 inch $=2\cdot5399772$ centimetres.

Area.

1 sq. inch $=6\cdot45$ sq. cm., nearly.
1 sq. foot $=929$ sq. cm., nearly.
1 sq. yard $=8361$ sq. cm., nearly.
1 sq. mile $=2\cdot59\times10^{10}$ sq. cm., nearly.

Volume.

1 cub. inch $=16\cdot39$ cub. cm., nearly.
1 cub. foot $=28316$ cub. cm., nearly.
1 cub. yard $=764535$ cub. cm., nearly.
1 gallon $=4541$ cub. cm., nearly.

Mass.

1 grain $=\cdot0648$ gramme, nearly.
1 oz. avoir. $=28\cdot35$ gramme, nearly.
1 lb. avoir. $=453\cdot6$ gramme, nearly.
1 ton $=1\cdot016\times10^{6}$ gramme, nearly
More accurately, 1 lb. avoir. $=453\cdot59265$ gm.

Velocity.

1 mile per hour $=44\cdot704$ cm. per sec.
1 kilometre per hour $=27\cdot7$ cm. per sec.

Density.

1 lb. per cub. foot $=\cdot016019$ gm. per cub. cm.
$62\cdot4$ lbs. per cub. ft. $=1$ gm. per cub. cm.

FORCE (assuming $g=981$). (See p. 48.)

Weight of 1 grain	$=63 \cdot 57$ dynes, nearly.
,, 1 oz. avoir.	$=2 \cdot 78 \times 10^4$ dynes, nearly.
,, 1 lb. avoir.	$=4 \cdot 45 \times 10^5$ dynes, nearly.
,, 1 ton	$=9 \cdot 97 \times 10^8$ dynes, nearly.
,, 1 gramme	$=981$ dynes, nearly.
,, 1 kilogramme	$=9 \cdot 81 \times 10^5$ dynes, nearly.

WORK (assuming $g=981$). (See p. 48.)

1 foot-pound	$=1 \cdot 356 \times 10^7$ ergs, nearly.
1 kilogrammetre	$=9 \cdot 81 \times 10^7$ ergs, nearly.
Work in a second by one theoretical "horse."	$=7 \cdot 46 \times 10^9$ ergs, nearly.

STRESS (assuming $g=981$).

1 lb. per sq. ft.	$=479$ dynes per sq. cm., nearly.
1 lb. per sq. inch	$=6 \cdot 9 \times 10^4$ dynes per sq. cm., nearly.
1 kilog. per sq. cm.	$=9 \cdot 81 \times 10^5$ dynes per sq. cm., nearly.
760 mm. of mercury at 0°C.	$=1 \cdot 014 \times 10^6$ dynes per sq. cm., nearly.
30 inches of mercury at 0° C.	$=1 \cdot 0163 \times 10^6$ dynes per sq. cm., nearly.
1 inch of mercury at 0° C.	$=3 \cdot 388 \times 10^4$ dynes per sq. cm., nearly.

TABLE OF CONSTANTS.

The velocity acquired in falling for one second in vacuo, in any part of Great Britain, is about 32·2 feet per second, or 9·81 metres per second.

The pressure of one atmosphere, or 760 millimetres (29·922 inches) of mercury, is 1·033 kilogramme per sq. centimetre, or 14·73 lbs. per square inch.

The weight of a litre of dry air, at this pressure (at Paris) and 0° C., is 1·293 gramme.

The weight of a cubic centimetre of water is about 1 gramme.

The weight of a cubic foot of water is about 62·4 lbs.

The equivalent of a unit of heat, in gravitation units of energy, is—

772 for the foot and Fahrenheit degree.
1390 for the foot and Centigrade degree.
424 for the metre and Centigrade degree.
42400 for the centimetre and Centigrade degree.

In absolute units of energy, the equivalent is—

41·6 millions for the centimetre and Centigrade degree;

or 1 gramme-degree is equivalent to 41·6 million ergs.

ACOUSTICS.

CHAPTER LXII.

PRODUCTION AND PROPAGATION OF SOUND.

866. Sound is a Vibration.—Sound, as directly known to us by the sense of hearing, is an impression of a peculiar character, very broadly distinguished from the impressions received through the rest of our senses, and admitting of great variety in its modifications. The attempt to explain the physiological actions which constitute hearing forms no part of our present design. The business of physics is rather to treat of those external actions which constitute sound, considered as an objective existence external to the ear of the percipient.

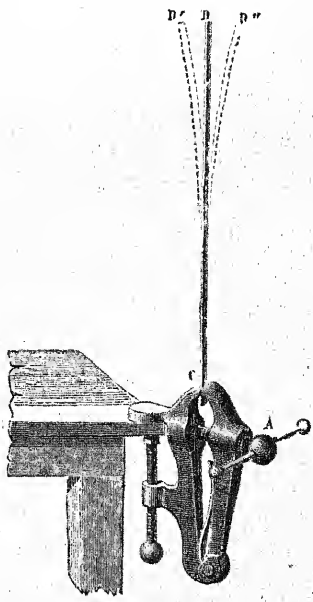


Fig. 592.—Vibration of Straight Spring.

It can be shown, by a variety of experiments, that sound is the result of vibratory movement. Suppose, for example, we fix one end C of a straight spring CD (Fig. 592) in a vice A, then draw the other end D aside into the position D′, and let it go. In virtue of its elasticity (§ 126), the spring will return to its original position; but the kinetic energy which it acquires in returning is sufficient to carry it to a nearly equal distance on the other side; and it thus swings alternately from one side to the other through distances very gradually diminishing, until at last it comes to rest. Such movement is called vibratory. The motion from D′ to D″, or from D″ to D′, is called a *single vibration*. The two together constitute a

double or *complete vibration;* and the time of executing a complete vibration is the *period* of vibration. The *amplitude* of vibration for any point in the spring is the distance of its middle position from one of its extreme positions. These terms have been already employed (§ 107) in connection with the movements of pendulums to which indeed the movements of vibrating springs bear an extremely close resemblance. The property of isochronism, which approximately characterizes the vibrations of the pendulum, also belongs to the spring, the approximation being usually so close that the period may practically be regarded as altogether independent of the amplitude.

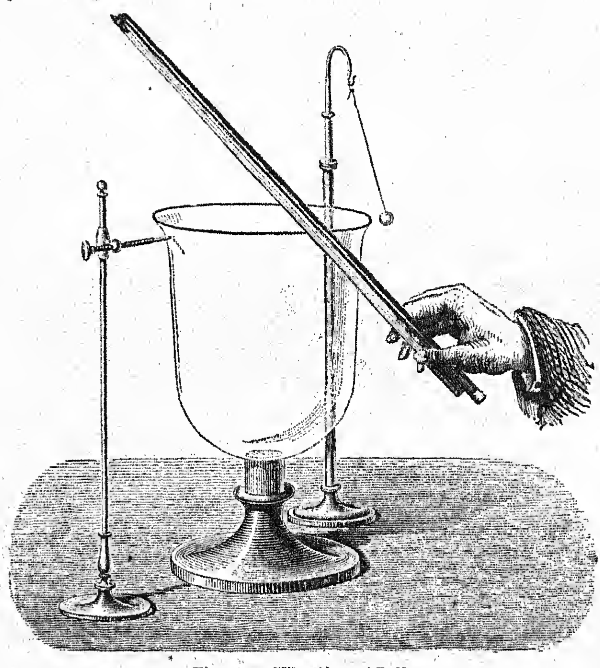


Fig. 593.—Vibration of Bell.

When the spring is long, the extent of its movements may generally be perceived by the eye. In consequence of the persistence of impressions, we see the spring in all its positions at once; and the edges of the space moved over are more conspicuous than the central parts, because the motion of the spring is slowest at its extreme positions.

As the spring is lowered in the vice, so as to shorten the vibrating portion of it, its movements become more rapid, and at the same time

more limited, until, when it is very short, the eye is unable to detect any sign of motion. But where sight fails us, hearing comes to our aid. As the vibrating part is shortened more and more, it emits a musical note, which continually rises in pitch; and this effect continues after the movements have become much too small to be visible.

It thus appears that a vibratory movement, if sufficiently rapid, may produce a sound. The following experiments afford additional illustration of this principle, and are samples of the evidence from which it is inferred that vibratory movement is essential to the production of sound.

Vibration of a Bell.—A point is fixed on a stand, in such a position as to be nearly in contact with a glass bell (Fig. 593). If a rosined fiddle-bow is then drawn over the edge of the bell, until a

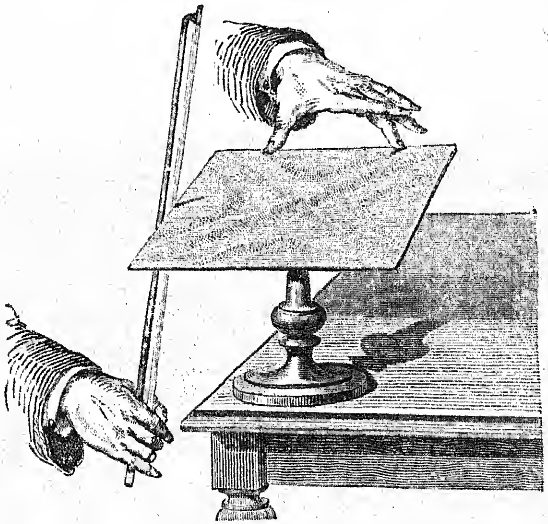


Fig. 594.—Vibration of Plate.

musical note is emitted, a series of taps are heard, due to the striking of the bell against the point. A pith-ball, hung by a thread, is driven out by the bell, and kept in oscillation as long as the sound continues. By lightly touching the bell, we may feel that it is vibrating; and if we press strongly, the vibration and the sound will both be stopped.

Vibration of a Plate.—Sand is strewn over the surface of a horizontal plate (Fig. 594), which is then made to vibrate by drawing a

bow over its edge. As soon as the plate begins to sound, the sand dances, leaves certain parts bare, and collects in definite lines, which are called *nodal lines.* These are, in fact, the lines which separate portions of the plate whose movements are in opposite directions. Their position changes whenever the plate changes its note.

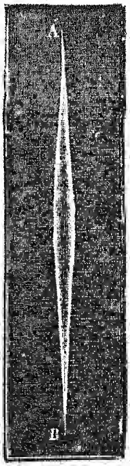


Fig. 595.
Vibration of String.

The vibratory condition of the plate is also manifested by another phenomenon, opposite—so to speak—to that just described. If very fine powder, such as lycopodium, be mixed with the sand, it will not move with the sand to the nodal lines, but will form little heaps in the centre of the vibrating segments; and these heaps will be in a state of violent agitation, with more or less of gyratory movement, as long as the plate is vibrating. This phenomenon, after long baffling explanation, was shown by Faraday to be due to indraughts of air, and ascending currents, brought about by the movements of the plate. In a moderately good vacuum, the lycopodium goes with the sand to the nodal lines.

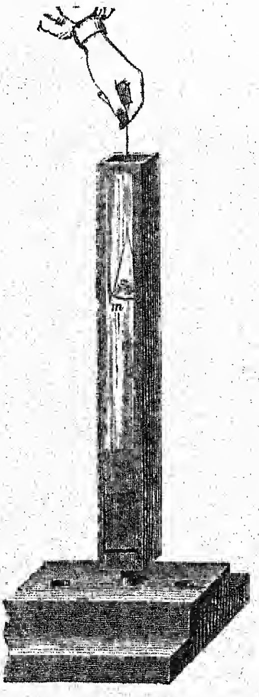


Fig. 596.—Vibration of Air.

Vibration of a String.—When a note is produced from a musical string or wire, its vibrations are often of sufficient amplitude to be detected by the eye. The string thus assumes the appearance of an elongated spindle (Fig. 595).

Vibration of the Air.—The sonorous body may sometimes be air, as in the case of organ-pipes, which we shall describe in a later chapter. It is easy to show by experiment that when a pipe *speaks*, the air within it is vibrating. Let one side of the tube be of glass, and let a small membrane m, stretched over a frame, be strewed with sand, and lowered into the pipe. The sand will be thrown into violent agitation, and the rattling of the grains, as they fall back on the membrane, is loud enough to be distinctly heard.

Singing Flames.—An experiment on the production of musical sound by flame, has long been known under the name of the *chemical harmonica.* An apperatus for the production of hydrogen gas (Fig. 597) is furnished with a tube, which tapers off nearly to a point at its upper end, where the gas issues and is lighted. When a tube, open at both ends, is held so as to surround the flame, a musical tone is heard, which varies with the dimensions of the tube, and often attains considerable power. The sound is due to the vibration of the air and products of combustion within the tube; and on observing the reflection of the flame in a mirror rotating about a vertical axis, it will be seen that the flame is alternately rising and falling, its successive images, as drawn out into a horizontal series by the rotation of the mirror, resembling a number of equidistant tongues of flame, with depressions between them. The experiment may also be performed with ordinary coal-gas.

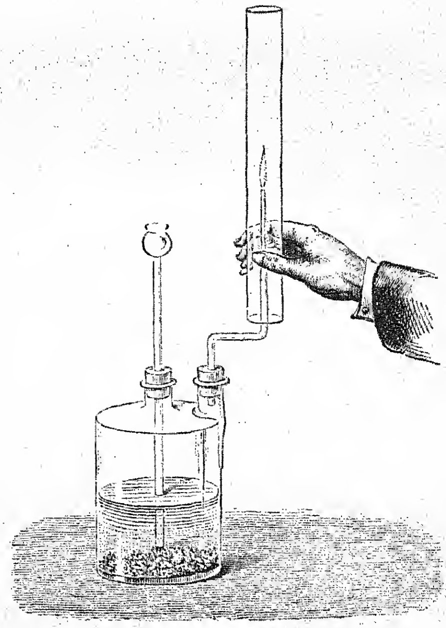


Fig. 597.—Chemical Harmonica.

Trevelyan Experiment.—A fire-shovel (Fig. 598) is heated, and balanced upon the edges of two sheets of lead fixed in a vice; it is then seen to execute a series of small oscillations—each end being alternately raised and depressed—and a sound is at the same time emitted. The oscillations are so small as to be scarcely perceptible in themselves; but they can be rendered very obvious by attaching to the shovel a small silvered mirror, on which a beam of light is directed. The reflected light can be made to form an image upon a screen, and this image is seen to be in a state of oscillation as long as the sound is heard.

The movements observed in this experiment are due to the sudden expansion of the cold lead. When the hot iron comes in contact with

it, a protuberance is instantly formed by dilatation, and the iron is thrown up. It then comes in contact with another portion of the

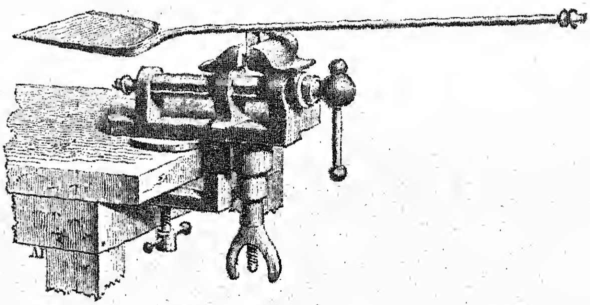


Fig. 598.—Trevelyan Experiment.

lead, where the same phenomenon is repeated while the first point cools. By alternate contacts and repulsions at the two points, the shovel is kept in a continual state of oscillation, and the regular succession of taps produces the sound.

The experiment is more usually performed with a special instrument invented by Trevelyan, and called a *rocker*, which, after being heated and laid upon a block of lead, rocks rapidly from side to side, and yields a loud note.

867. Distinctive Character of Musical Sound.—It is not easy to draw a sharp line of demarcation between musical sound and mere noise. The name of noise is usually given to any sound which seems unsuited to the requirements of music.

This unfitness may arise from one or the other of two causes. Either,

1. The sound may be unpleasant from containing discordant elements which jar with one another, as when several consecutive keys on a piano are put down together. Or,

2. It may consist of a confused succession of sounds, the changes being so rapid that the ear is unable to identify any particular note. This kind of noise may be illustrated by sliding the finger along a violin-string, while the bow is applied.

All sounds may be resolved into combinations of elementary musical tones occurring simultaneously and in succession. Hence the study of musical sounds must necessarily form the basis of acoustics.

Every sound which is recognized as musical is characterized by what may be called smoothness, evenness, or regularity; and the physical cause of this regularity is to be found in the accurate

periodicity of the vibratory movements which produce the sound. By *periodicity* we mean the recurrence of precisely similar states at equal intervals of time, so that the movements exactly repeat themselves; and the time which elapses between two successive recurrences of the same state is called the *period* of the movements.

Practically, musical and unmusical sounds often shade insensibly into one another. The tones of every musical instrument are accompanied by more or less of unmusical noise. The sounds of bells and drums have a sort of intermediate character; and the confused assemblage of sounds which is heard in the streets of a city blends at a distance into an agreeable hum.

868. Vehicle of Sound.—The origin of sound is always to be found in the vibratory movements of a sonorous body; but these vibratory movements cannot bring about the sensation of hearing unless there be a medium to transmit them to the auditory apparatus. This medium may be either solid, liquid, or gaseous, but it is necessary that it be elastic. A body vibrating in an absolute vacuum, or in a medium utterly destitute of elasticity, would fail to excite our sensations of hearing. This assertion is justified by the following experiments:—

1. Under the receiver of an air-pump is placed a clock-work arrangement for producing a number of strokes on a bell.

It is placed on a thick cushion of felt, or other inelastic material, and the air in the receiver is exhausted as completely as possible. If the clock-work is then started by means of the handle g, the hammer will be seen to strike the bell, but the sound will be scarcely audible. If hydrogen be introduced into the vacuum, and the receiver be again exhausted, the sound will be much more completely extinguished, being heard with difficulty even when the ear is placed in contact with the receiver. Hence it may fairly be concluded that if the receiver could be perfectly exhausted, and a perfectly inelastic support could be found for the bell, no sound at all would be emitted.

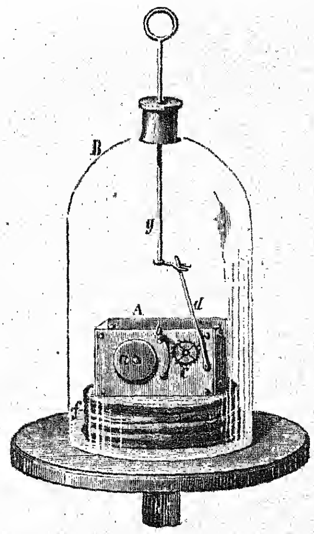


Fig. 599.—Sound in Exhausted Receiver.

2. The experiment may be varied by using a glass globe, furnished with a stop-cock, and having a little bell suspended within it by a thread. If the globe is exhausted of air, the sound of the bell will be scarcely audible. The globe may be filled with any kind of gas, or with vapour either saturated or non-saturated, and it will thus be found that all these bodies transmit sound.

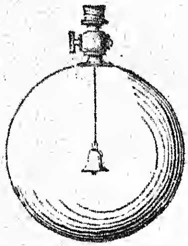


Fig. 600.
Globe with Stop-cock.

Sound is also transmitted through liquids, as may easily be proved by direct experiment. Experiment, however, is scarcely necessary for the establishment of the fact, seeing that fishes are provided with auditory apparatus, and have often an acute sense of hearing.

As to solids, some well-known facts prove that they transmit sound very perfectly. For example, light taps with the head of a pin on one end of a wooden beam, are distinctly heard by a person with his ear applied to the other end, though they cannot be heard at the same distance through the air. This property is sometimes employed as a test of the soundness of a beam, for the experiment will not succeed if the intervening wood is rotten, rotten wood being very inelastic.

The *stethoscope* is an example of the transmission of sound through solids. It is a cylinder of wood, with an enlargement at each end, and a perforation in its axis. One end is pressed against the chest of the patient, while the observer applies his ear to the other. He is thus enabled to hear the sounds produced by various internal actions, such as the beating of the heart and the passage of the air through the tubes of the lungs. Even simple *auscultation*, in which the ear is applied directly to the surface of the body, implies the transmission of sound through the walls of the chest.

By applying the ear to the ground, remote sounds can often be much more distinctly heard; and it is stated that savages can in this way obtain much information respecting approaching bodies of enemies.

We are entitled then to assert that *sound, as it affects our organs of hearing, is an effect which is propagated, from a vibrating body, through an elastic and ponderable medium.*

869. Mode of Propagation of Sound.—We will now endeavour to explain the action by which sound is propagated.

Let there be a plate *a* vibrating opposite the end of a long tube, and let us consider what happens during the passage of the plate

from its most backward position a'', to its most advanced position a'. This movement of the plate may be divided in imagination into a number of successive parts, each of which is communicated to the layer of air close in front of it, which is thus compressed, and, in its

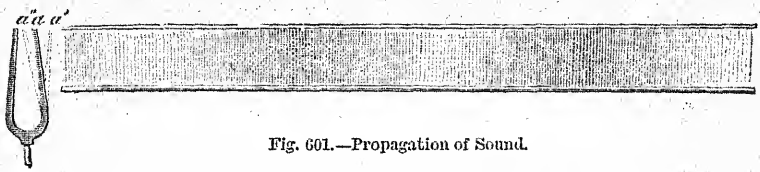


Fig. 601.—Propagation of Sound.

endeavour to recover from this compression, reacts upon the next layer, which is thus in its turn compressed. The compression is thus passed on from layer to layer through the whole tube, much in the same way as, when a number of ivory balls are laid in a row, if the first receives an impulse which drives it against the second, each ball will strike against its successor and be brought to rest.

The compression is thus passed on from layer to layer through the tube, and is succeeded by a rarefaction corresponding to the backward movement of the plate from a' to a''. As the plate goes on vibrating, these compressions and rarefactions continue to be propagated through the tube in alternate succession. The greatest compression in the layer immediately in front of the plate, occurs when the plate is at its middle position in its forward movement, and the greatest rarefaction occurs when it is in the same position in its backward movement. These are also the instants at which the plate is moving most rapidly.[1] When the plate is in its most advanced position, the layer of air next to it, A (Fig. 602) will be in its natural state, and another layer at A_1, half a wave-length further on, will also be in its natural state, the pulse having travelled from A to A_1, while the plate was moving from a'' to a'. At intervening points between A and A_1, the layers will have various amounts of compression corresponding to the different positions of the plate in its forward movement. The greatest compression is at C, a quarter of a wave-length in advance of A, having travelled over this distance while the plate was advancing from a to a'. The compressions at D and D_1 repre-

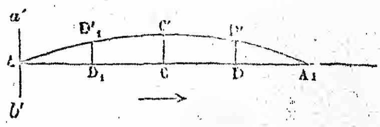


Fig. 602.—Graphical Representation.

[1] See § 870, also Note A at the end of this chapter.

sent those which existed immediately in front of the plate when it had advanced respectively one-fourth and three-fourths of the distance from a'' to a', and the curve A C′ A_1 is the graphical representation both of condensation and velocity for all points in the air between A and A_1.

If the plate ceased vibrating, the condition of things now existing in the portion of air AA_1 would be transferred to successive portions of air in the tube, and the curve A C′A_1 would, as it were, slide onward through the tube with the velocity of sound, which is about 1100 feet per second. But the plate, instead of remaining permanently at a', executes a backward movement, and produces rarefactions and retrograde velocities, which are propagated onwards in the same manner as the condensations and forward velocities. A complete wave of the undulation is accordingly represented by the curve A E′A_1 C′A_2 (Fig. 603), the portions of the curve below the line of

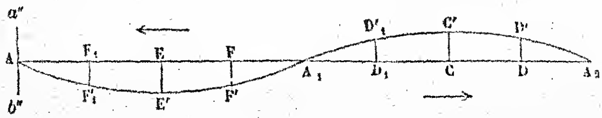


Fig. 603.—Graphical Representation of Complete Wave.

abscissas being intended to represent rarefactions and retrograde velocities. If we suppose the vibrating plate to be rigidly connected with a piston which works air-tight in the tube, the velocities of the particles of air in the different points of a wave-length will be identical with the velocities of the piston at the different parts of its motion.

The wave-length A A_2 is the distance that the pulse has travelled while the vibrating plate was moving from its most backward to its most advanced position, and back again. During this time, which is called the *period* of the vibrations, each particle of air goes through its complete cycle of changes, both as regards motion and density. The period of vibration of any particle is thus identical with that of the vibrating plate, and is the same as the time occupied by the waves in travelling a wave-length. Thus, if the plate be one leg of a common A tuning-fork, making 435 complete vibrations per second, the period will be $\frac{1}{435}$th of a second, and the undulation will travel in this time a distance of $\frac{1100}{435}$ feet, or 2 feet 6 inches, which is therefore the wave-length in air for this note. If the plate continues to vibrate in a uniform manner, there will be a continual series of equal

and similar waves running along the tube with the velocity of sound. Such a succession of waves constitutes an undulation. Each wave consists of a condensed portion, and a rarefied portion, which are distinguished from each other in Fig. 601 by different tints, the dark shading being intended to represent condensation.

870. Nature of Undulations.—The possibility of condensations and rarefactions being propagated continually in one direction, while each particle of air simply moves backwards and forwards about its original position, is illustrated by Fig. 604, which represents, in an

```
      A B    C          D          E     F a
         AB  C       .D          E       F   a
           A B C    D          E         F    a
            A  BC  D        E          F      a
            A    B C D     E        F         a
         A       B  CD   E       F          a
      A          B     C D E    F       a
   A           B       C  DE   F      a
 A          B          C    D E F   a
A         B          C      D   EF  a
 A      B        C         D    E F a
   A   B       C         D     E  F a
     A B     C        D        E    F a
```

Fig. 604.—Longitudinal Vibration.

exaggerated form, the successive phases of an undulation propagated through 7 particles A B C D E F a originally equidistant, the distance from the first to the last being one wave-length of the undulation. The diagram is composed of thirteen horizontal rows, the first and last being precisely alike. The successive rows represent the positions of the particles at successive times, the interval of time from each row to the next being $\frac{1}{12}$th of the period of the undulation.

In the first row A and a are centres of condensation, and D is a centre of rarefaction. In the third row B is a centre of condensation, and E a centre of rarefaction. In the fifth row the condensation and rarefaction have advanced by one more letter, and so on through the whole series, the initial state of things being

reproduced when each of these centres has advanced through a wave-length, so that the thirteenth row is merely a repetition of the first.

The velocities of the particles can be estimated by the comparison of successive rows. It is thus seen that the greatest forward velocity is at the centres of condensation, and the greatest backward velocity at the centres of rarefaction. Each particle has its greatest velocities, and greatest condensation and rarefaction, in passing through its mean position, and comes for an instant to rest in its positions of greatest displacement, which are also positions of mean density.

The distance between A and a remains invariable, being always a wave-length, and these two particles always agree in phase. Any other two particles represented in the diagram are always in different phases, and the phases of A and D, or B and E, or C and F, are always opposite; for example, when A is moving forwards with the maximum velocity, D is moving backwards with the same velocity.

The vibrations of the particles, in an undulation of this kind, are called *longitudinal;* and it is by such vibrations that sound is propagated through air. Fig. 605 illustrates the manner in which an undulation may be propagated by means of *transverse* vibrations, that is to say, by vibrations executed in a direction perpendicular to that in which the undulation advances. Thirteen particles A B C D E F G H I J K L a are represented in the positions which they occupy at successive times, whose interval is one-sixth of a period. At the instant first considered, D and J are the particles which are furthest displaced. At the end of the first interval, the wave has advanced two letters, so that F and L are now the furthest displaced. At the end of the next interval, the wave has advanced two letters further, and so on, the state of things at the end of the six intervals, or of one complete period, being the same as at the beginning, so that the seventh line is merely a repetition of the first. Some examples of this kind of wave-motion will be mentioned in later chapters.

871. Propagation in an Open Space.—When a sonorous disturbance occurs in the midst of an open body of air, the undulations to which it gives rise run out in all directions from the source. If the disturbance is symmetrical about a centre, the waves will be spherical; but this case is exceptional. A disturbance usually produces condensation on one side, at the same instant that it produces rarefaction on another. This is the case, for example, with a vibrating

plate, since, when it is moving towards one side, it is moving away from the other. These inequalities which exist in the neighbourhood of the sonorous body, have, however, a tendency to become less marked, and ultimately to disappear, as the distance is increased. Fig. 606 represents a diametral section of a series of spherical waves. Their mode of propagation has some analogy to that of the circular

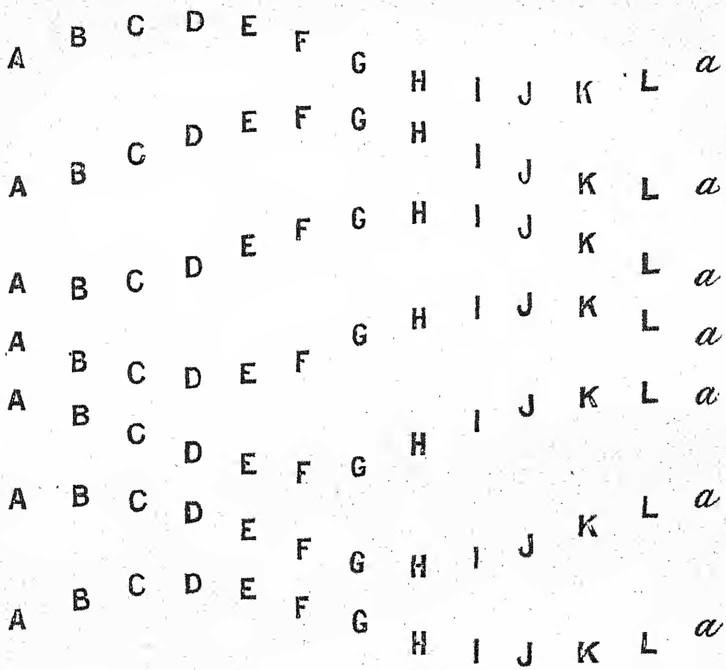


Fig. 605.—Transverse Vibration.

waves produced on water by dropping a stone into it; but the particles which form the waves of water rise and fall, whereas those which form sonorous waves merely advance and retreat, their lines of motion being always coincident with the directions along which the sound travels. In both cases it is important to remark that *the undulation does not involve a movement of transference.* Thus, when the surface of a liquid is traversed by waves, bodies floating on it rise and fall, but are not carried onward. This property is characteristic of undulations generally. *An undulation may be defined as a system of movements in which the several particles move to and fro, or round and round, about definite points, in such a*

manner as to produce the continued onward transmission of a condition, or series of conditions.

There is one important difference between the propagation of sound in a uniform tube and in an open space. In the former case, the layers of air corresponding to successive wave-lengths are of equal

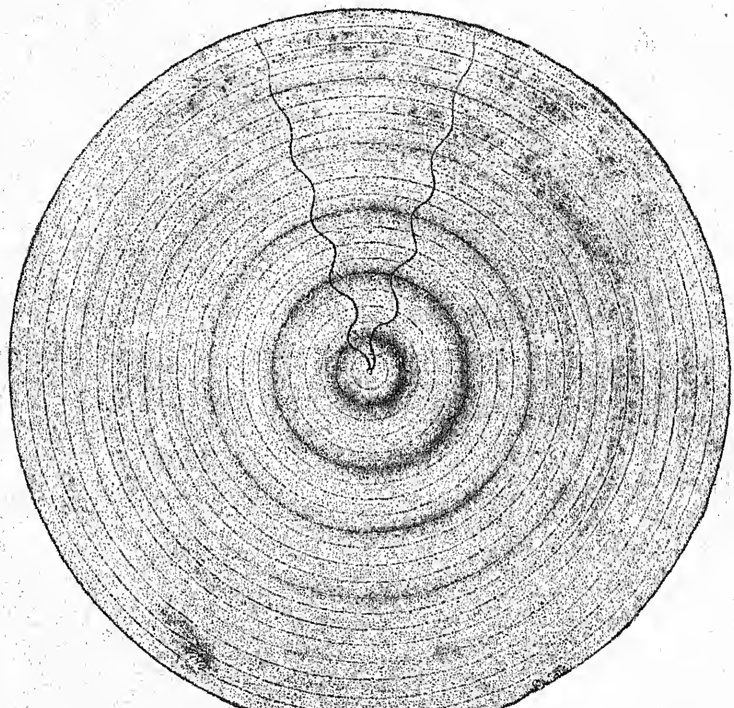


Fig. 606.—Propagation in Open Space.

mass, and their movements are precisely alike, except in so far as they are interfered with by friction. Hence sound is transmitted through tubes to great distances with but little loss of intensity, especially if the tubes are large.[1]

The same principle is illustrated by the ease with which a scratch

[1] Regnault, in his experiments on the velocity of sound, found that in a conduit ·108 of a metre in diameter, the report of a pistol charged with a gramme of powder ceased to be heard at the distance of 1150 metres. In a conduit of ·3^{m}, the distance was 3810^{m}. In the great conduit of the St. Michel sewer, of 1^{m}·10, the sound was made by successive reflections to traverse a distance of 10,000 metres without becoming inaudible.—*D.*

on a log of wood is heard at the far end, the substance of the log acting like the body of air within a tube.

In an open space, each successive layer has to impart its own condition to a larger layer; hence there is a continual diminution of amplitude in the vibrations as the distance from the source increases. This involves a continual decrease of loudness. An undulation involves the onward transference of energy; and the amount of energy which traverses, in unit time, any closed surface described about the source, must be equal to the energy which the source emits in unit time. Hence, by the reasoning which we employed in the case of radiant heat (§ 465), it follows that the intensity of sonorous energy diminishes according to the law of inverse squares.

The energy of a particle executing simple vibrations in obedience to forces of elasticity, varies as the square of the amplitude of its excursions; for, if the amplitude be doubled, the distance worked through, and the mean working force, are both doubled, and thus the work which the elastic forces do during the movement from either extreme position to the centre is quadrupled. This work is equal to the energy of the particle in any part of its course. At the extreme positions it is all in the shape of potential energy; in the middle position it is all in the shape of kinetic energy; and at intermediate points it is partly in one of these forms, and partly in the other.

It can be shown that exactly half the energy of a complete wave is kinetic, the other half being potential.

872. Dissipation of Sonorous Energy.—The reasoning by which we have endeavoured to establish the law of inverse squares, assumes that onward propagation involves no loss of sonorous energy. This assumption is not rigorously true, inasmuch as vibration implies friction, and friction implies the generation of heat, at the expense of the energy which produces the vibrations. Sonorous energy must therefore diminish with distance somewhat more rapidly than according to the law of inverse squares. All sound, in becoming extinct, becomes converted into heat.

This conversion is greatly promoted by defect of homogeneity in the medium of propagation. In a fog, or a snow-storm, the liquid or solid particles present in the air produce innumerable reflections, in each of which a little sonorous energy is converted into heat.

873. Velocity of Sound in Air.—The propagation of sound through an elastic medium is not instantaneous, but occupies a very sensible time in traversing a moderate distance. For example, the flash of a gun at the distance of a few hundred yards is seen some time before the report is heard. The interval between the two impressions may be regarded as representing the time required for the propagation of the sound across the intervening distance, for the time occupied by the propagation of light across so small a distance is inappreciable.

It is by experiments of this kind that the velocity of sound in air has been most accurately determined. Among the best determinations may be mentioned that of Lacaille, and other members of a commission appointed by the French Academy in 1738; that of Arago, Bouvard, and other members of the Bureau de Longitudes in 1822; and that of Moll, Vanbeek, and Kuytenbrouwer in Holland, in the same year. All these determinations were obtained by firing cannon at two stations, several miles distant from each other, and noting, at each station, the interval between seeing the flash and hearing the sound of the guns fired at the other. If guns were fired only at one station, the determination would be vitiated by the effect of wind blowing either with or against the sound. The error from this cause is nearly eliminated by firing the guns alternately at the two stations, and still more completely by firing them simultaneously This last plan was adopted by the Dutch observers, the distance of the two stations in their case being about nine miles. Regnault has quite recently repeated the investigation, taking advantage of the important aid afforded by modern electrical methods for registering the times of observed phenomena. All the most careful determinations agree very closely among themselves, and show that the velocity of sound through air at 0° C. is about 332 metres, or 1090 feet per second.[1] The velocity increases with the temperature, being proportional to *the square root of the absolute temperature by air thermometer* (§ 325). If t denote the ordinary Centigrade tempera-

[1] A recent determination by Mr. Stone at the Cape of Good Hope is worthy of note as being based on the comparison of observations made through the sense of hearing alone. It had previously been suggested that the two senses of sight and hearing, which are concerned in observing the flash and report of a cannon, might not be equally prompt in receiving impressions (Airy on *Sound*, p. 131). Mr. Stone accordingly placed two observers —one near a cannon, and the other at about three miles distance; each of whom on hearing the report, gave a signal through an electric telegraph. The result obtained was in precise agreement with that stated in the text.

ture, and a the coefficient of expansion ·00366, the velocity of sound through air at any temperature is given by the formula

$$332 \sqrt{1+a t} \text{ in metres per second, or}$$
$$1090 \sqrt{1+a t} \text{ in feet per second.}$$

The actual velocity of sound from place to place on the earth's surface is found by compounding this velocity with the velocity of the wind.

There is some reason, both from theory and experiment, for believing that very loud sounds travel rather faster than sounds of moderate intensity.

874. Theoretical Computation of Velocity.—By applying the principles of dynamics to the propagation of undulations,[1] it is computed that the velocity of sound through air must be given by the formula

$$v=\sqrt{\frac{E}{D}}. \qquad (1)$$

D denoting the density of the air, and E its coefficient of elasticity, as measured by the quotient of pressure applied by compression produced.

Let P denote the pressure of the air in units of force per unit of area; then, if the temperature be kept constant during compression, a small additional pressure p will, by Boyle's law, produce a compression equal to $\frac{p}{P}$, and the value of E, being the quotient of p by this quantity, will be simply P.

On the other hand, if no heat is allowed either to enter or escape, the temperature of the air will be raised by compression, and additional resistance will thus be encountered. In this case, as shown in § 500, the coefficient of elasticity will be Pk, k denoting the ratio of the two specific heats, which for air and simple gases is about 1·41.

It thus appears that the velocity of sound in air cannot be less than $\sqrt{\frac{P}{D}}$ nor greater than $\sqrt{1{\cdot}41\,\frac{P}{D}}$. Its actual velocity, as determined by observation, is identical, or practically identical, with the latter of these limiting values. Hence we must infer that the compressions and extensions which the particles of air undergo in transmitting sound are of too brief duration to allow of any sensible transference of heat from particle to particle.

This conclusion is confirmed by another argument due to Professor

[1] See note B at the end of this chapter.

Stokes. If the inequalities of temperature due to compression and expansion were to any sensible degree smoothed down by conduction and radiation, this smoothing down would diminish the amount of energy available for wave-propagation, and would lead to a falling off in intensity incomparably more rapid than that due to the law of inverse squares.

875. Numerical Calculation.—The following is the actual process of calculation for perfectly dry air at 0° C., the centimetre, gramme, and second being taken as the units of length, mass, and time.

The density of dry air at 0°, under the pressure of 1033 grammes per square centimetre, at Paris, is ·001293 of a gramme per cubic centimetre. But the gravitating force of a gramme at Paris is 981 dynes (§ 91). The density ·001293 therefore corresponds to a pressure of 1033×981 dynes per sq. cm.; and the expression for the velocity in centimetres per second is

$$v = \sqrt{1{\cdot}41\,\frac{P}{D}} = \sqrt{1{\cdot}41\,\frac{1033 \times 981}{{\cdot}001293}} = 33240 \text{ nearly};$$

that is, 332·4 metres per second, or 1093 feet per second. 107.

876. Effects of Pressure, Temperature, and Moisture.—The velocity of sound is independent of the height of the barometer, since changes of this element (at constant temperature) affect P and D in the same direction, and to the same extent.

For a given density, if P_0 denote the pressure at 0°, and α the coefficient of expansion of air, the pressure at t° Centigrade is $P_0\,(1 + \alpha\, t)$, the value of α being about $\frac{1}{273}$.

Hence, if the velocity at 0° be 1090 feet per second, the velocity at t° will be $1090\,\sqrt{1 + \frac{t}{273}}$. At the temperature 50° F. or 10° C., which is approximately the mean annual temperature of this country, the value of this expression is about 1110, and at 86° F. or 30° C. it is about 1148. The increase of velocity is thus about a foot per second for each degree Fahrenheit.

The humidity of air has some influence on the velocity of sound, inasmuch as aqueous vapour is lighter than air; but the effect is comparatively trifling, at least in temperate climates. At the temperature 50° F., air saturated with moisture is less dense than dry air by about 1 part in 220, and the consequent increase of velocity cannot be greater than about 1 part in 440, which will be between 2 and 3 feet per second. The increase should, in fact, be somewhat

less than this, inasmuch as the value of k (the ratio of the two specific heats) appears to be only 1·31 for aqueous vapour.[1]

877. Newton's Theory, and Laplace's Modification.—The earliest theoretical investigation of the velocity of sound was that given by Newton in the *Principia* (book 2, section 8). It proceeds on the tacit assumption that no changes of temperature are produced by the compressions and extensions which enter into the constitution of a sonorous undulation; and the result obtained by Newton is equivalent to the formula

$$v=\sqrt{\frac{P}{D}};$$

or since (§ 210) $\frac{P}{D}=g\,H$, where H denotes the *height of a homogeneous atmosphere*, and the velocity acquired in falling through any height s is $\sqrt{2\,g\,s}$, the velocity of sound in air is, according to Newton, the same as *the velocity which would be acquired by falling in vacuo through half the height of a homogeneous atmosphere.* This, in fact, is the form in which Newton states his result.[2]

Newton himself was quite aware that the value thus computed theoretically was too small, and he throws out a conjecture as to the cause of the discrepancy; but the true cause was first pointed out by Laplace, as depending upon increase of temperature produced by compression, and decrease of temperature produced by expansion.

878. Velocity in Gases generally.—The same principles which apply to air apply to gases generally; and since for all simple gases the ratio of the two specific heats is 1·41, the velocity of sound in any simple gas is $\sqrt{1{\cdot}41\,\frac{P}{D}}$, D denoting its absolute density at the pressure P. Comparing two gases at the same pressure, we see that the velocities of sound in them will be inversely as the square roots of their absolute densities; and this will be true whether the temperatures of the two gases are the same or different.

879. Velocity of Sound in Liquids.—The velocity of sound in water was measured by Colladon, in 1826, at the Lake of Geneva. Two boats were moored at a distance of 13,500 metres (between 8 and 9 miles). One of them carried a bell, weighing about 140 lbs., immersed in the lake. Its hammer was moved by an external lever, so arranged as to ignite a small quantity of gunpowder at the instant

[1] Rankine on the *Steam Engine*, p. 320.

[2] Newton's investigation relates only to *simple* waves; but if these have all the same velocity (as Newton shows), this must also be the velocity of the complex wave which they compose. Hence the restriction is only apparent.

of striking the bell. An observer in the other boat was enabled to hear the sound by applying his ear to the extremity of a trumpet-shaped tube (Fig. 607), having its lower end covered with a membrane and facing towards the direction from which the sound proceeded. By noting the interval between seeing the flash and hearing the sound, the velocity with which the sound travelled through the water was determined. The velocity thus computed was 1435 metres per second, and the temperature of the water was 8°·1 C.

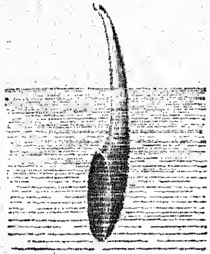


Fig. 607.

Formula (1) of § 874 holds for liquids as well as for gases.

The resistance of water to compression is about $2{\cdot}1 \times 10^{10}$ dynes per sq. cm., and the correcting factor for the heat of compression, as calculated by § 505, is 1·0012, which may be taken as unity. The density is also unity. Hence we have

$$v = \sqrt{\frac{E}{D}} = \sqrt{(2{\cdot}1 \times 10^{10})} = 144914 \text{ cm. per sec.},$$

that is, about 1449 metres per second; which agrees sufficiently well with the experimental determination.

Wertheim has measured the velocity of sound in some liquids by an indirect method, which will be explained in a later chapter. He finds it to be 1160 metres per second in ether and alcohol, and 1900 in a solution of chloride of calcium.

880. Velocity of Sound in Solids.—The velocity of sound in cast-iron was determined by Biot and Martin by means of a connected series of water-pipes, forming a conduit of a total length of 951 metres. One end of the conduit was struck with a hammer, and an observer at the other end heard two sounds, the first transmitted by the metal, and the second by the air, the interval between them being 2·5 seconds. Now the time required for travelling this distance through air, at the temperature of the experiment (11° C.), is 2·8 seconds. The time of transmission through the metal was therefore ·3 of a second, which is at the rate of 3170 metres per second. It is, however, to be remarked, that the transmitting body was not a continuous mass of iron, but a series of 376 pipes, connected together by collars of lead and tarred cloth, which must have considerably delayed the transmission of the sound. But in spite of this, the velocity is about nine times as great as in air.

Wertheim, by the indirect methods above alluded to, measured the velocity of sound in a number of solids, with the following results, the velocity in air being taken as the unit of velocity:—

Lead,	3·974 to 4·120	Steel,	14·361 to 15·108
Tin,	7·338 to 7·480	Iron,	15·108
Gold,	5·603 to 6·424	Brass,	10·224
Silver,	7·903 to 8·057	Glass,	14·956 to 16·759
Zinc,	9·863 to 11·009	Flint Glass,	11·890 to 12·220
Copper,	11·167	Oak,	9·902 to 12·02
Platinum,	7·823 to 8·467	Fir,	12·49 to 17·26

881. Theoretical Computation.—The formula $\sqrt{\frac{E}{D}}$ serves for solids as well as for liquids and gases; but as solids can be subjected to many different kinds of strain, whereas liquids and gases can be subjected to only one, we may have different values of E, and different velocities of transmission of pulses for the same solid. This is true even in the case of a solid whose properties are alike in all directions (called an *isotropic* solid); but the great majority of solids are very far from fulfilling this condition, and transmit sound more rapidly in some directions than in others.

When the sound is propagated by alternate compressions and extensions running along a substance which is not prevented from extending and contracting laterally, the elasticity E becomes identical[1] with Young's modulus (§ 128). On the other hand, if uniform spherical waves of alternate compression and extension spread outwards, symmetrically, from a point in the centre of an infinite solid, lateral extension and contraction will be prevented by the symmetry of the action. The effective elasticity is, in this case, greater than Young's modulus, and the velocity of sound will be increased accordingly.

By the table on p. 79 the value of Young's modulus for copper is 120×10^{10}, and by the table on p. xii. the density of copper is about 8·8. Hence, for the velocity of sound through a copper rod, in centimetres per second, we have

$$v = \sqrt{\frac{E}{D}} = \sqrt{\frac{120 \times 10^{10}}{8 \cdot 8}} = 369300 \text{ nearly,}$$

or 3693 metres per second.

This is about 11·1 times the velocity in air.

[1] Subject to a very small correction for heat of compression, which can be calculated by the formula of § 505. In the case of iron, the correcting factor is about 1·0023.

882. Reflection of Sound.—When sonorous waves meet a fixed obstacle they are reflected, and the two sets of waves—one direct, and the other reflected—are propagated just as if they came from two separate sources. If the reflecting surface is plane, waves di-

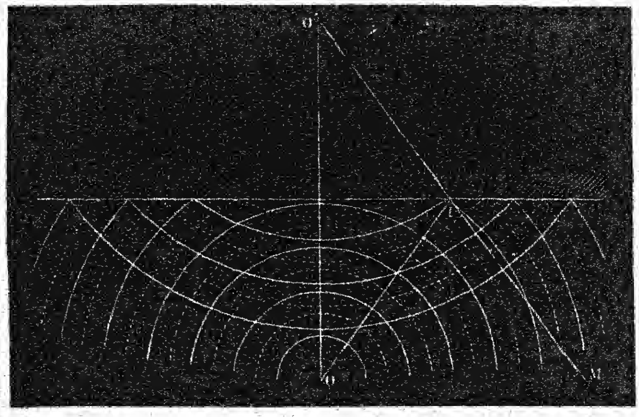


Fig. 608.—Reflection of Sound.

verging from any centre O (Fig. 608) in front of it are reflected so as to diverge from a centre O′ symmetrically situated behind it, and an ear at any point M in front hears the reflected sound as if it came from O′.

The direction from which a sound appears to the hearer to proceed is determined by the direction along which the sonorous pulses are propagated, and is always normal to the waves. A normal to a set of sound-waves may therefore conveniently be called a *ray* of sound. O I is a direct ray, and I M the corresponding reflected ray; and it is obvious, from the symmetrical position of the points O O′, that these two rays are equally inclined to the surface, or *the angles of incidence and reflection are equal.*

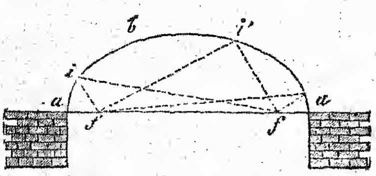


Fig. 609.—Reflection from Elliptic Roof.

883. Illustrations of Reflection of Sound.—The reflection of sonorous waves explains some well-known phenomena. If aba (Fig. 609) be an elliptic dome or arch, a sound emitted from either of the foci ff will be reflected from the elliptic surface in such a direction as to pass through the other focus. A sound emitted from either focus

may thus be distinctly heard at the other, even when quite inaudible at nearer points. This is a consequence of the property, that lines drawn to any point on an ellipse from the two foci are equally inclined to the curve.

The experiment of the conjugate mirrors (§ 468) is also applicable to sound. Let a watch be hung in the focus of one of them (Fig. 610),

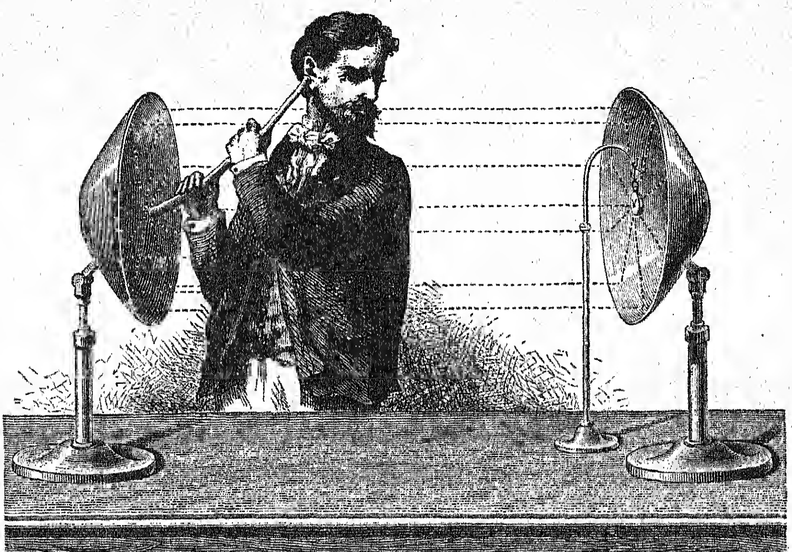


Fig. 610.—Reflection of Sound from Conjugate Mirrors.

and let a person hold his ear at the focus of the other; or still better, to avoid intercepting the sound before it falls on the second mirror, let him employ an ear-trumpet, holding its further end at the focus. He will distinctly hear the ticking, even when the mirrors are many yards apart.[1]

884. Echo.—Echo is the most familiar instance of the reflection of sound. In order to hear the echo of one's own voice, there must be a distant body capable of reflecting sound directly back, and the number of syllables that an echo will repeat is proportional to the

[1] Sondhaus has shown that sound, like light, is capable of being *refracted*. A spherical balloon of collodion, filled with carbonic acid gas, acts as a sound-lens. If a watch be hung at some distance from it on one side, an ear held at the conjugate focus on the other side will hear the ticking. See also a later section on "Curved Rays of Sound" in the chapter on the "Wave Theory of Light."

distance of this obstacle. The sounds reflected to the speaker have travelled first over the distance O A (Fig. 611) from him to the reflecting body, and then back from A to O. Supposing five syllables to be pronounced in a second, and taking the velocity of sound as 1100 feet per second, a distance of 550 feet from the speaker to the reflecting body would enable the speaker to complete the fifth syllable before the return of the first; this is at the rate of 110 feet

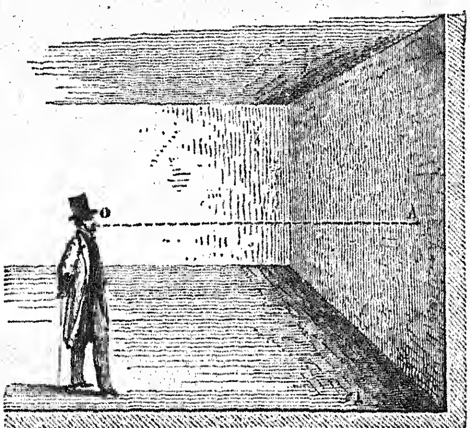


Fig. 611.—Echo.

per syllable. At distances less than about 100 feet there is not time for the distinct reflection of a single syllable; but the reflected sound mingles with the voice of the speaker. This is particularly observable under vaulted roofs.

Multiple echoes are not uncommon. They are due, in some cases, to independent reflections from obstacles at different distances; in others, to reflections of reflections. A position exactly midway between two parallel walls, at a sufficient distance apart, is favourable for the observance of this latter phenomenon. One of the most frequently cited instances of multiple echoes is that of the old palace of Simonetta, near Milan, which forms three sides of a quadrangle. According to Kircher, it repeats forty times.

885. Speaking and Hearing Trumpets.—The complete explanation of the action of these instruments presents considerable difficulty. The speaking-trumpet (Fig. 612) consists of a long tube (sometimes 6 feet long), slightly tapering towards the speaker, furnished at this end with a hollow mouth-piece, which nearly fits the lips, and at

the other with a funnel-shaped enlargement, called the *bell*, opening out to a width of about a foot. It is much used at sea, and is found very effectual in making the voice heard at a distance. The explanation usually given of its action is, that the slightly conical form of the long tube produces a series of reflections in directions more and more nearly parallel to the axis; but this explanation fails to account for the utility of the *bell*, which experience has shown to be considerable. It appears from a theoretical investigation by Lord Rayleigh that the speaking-trumpet causes a greater total quantity of sonorous energy to be produced from the same expenditure of breath.[1]

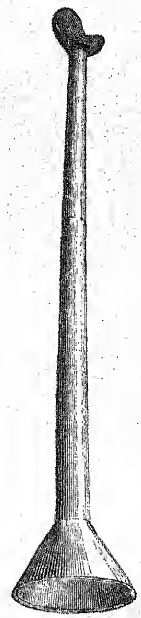


Fig. 612. Speaking-trumpet.

Ear-trumpets have various forms, as represented in Fig. 613; having little in common, except that the external opening or *bell* is much larger than the end which is introduced into the ear. Membranes of goldbeaters' skin are sometimes stretched across their interior, in the positions indicated by the dotted lines in Nos. 4 and 5. No. 6 consists simply of a bell with such a membrane stretched across its outer end, while its inner end communicates with the ear by an indian-rubber tube with an ivory end-piece. These light membranes are peculiarly susceptible of impression from aerial vibrations. In Regnault's experiments above cited, it was found that membranes were affected at distances greater than those at which sound was heard.

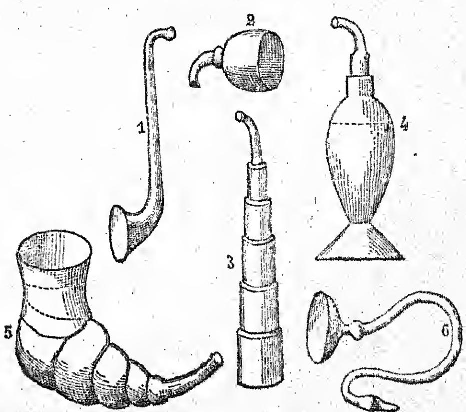


Fig. 613.—Ear-trumpets.

886. Interference of Sonorous Undulations.—When two systems of waves are traversing the same matter, the actual motion of each particle of the matter is the resultant of the motions due to each system separately. When these component motions are in the same direction the resultant is their

[1] *Theory of Sound*, vol. ii. p. 102.

sum; when they are in opposite directions it is their difference; and if they are equal, as well as opposite, it is zero. Very remarkable phenomena are thus produced when the two undulations have the same, or nearly the same wave-length; and the action which occurs in this case is called *interference.*

When two sonorous undulations of exactly equal wave-length and amplitude are traversing the same matter in the same direction, their phases must either be the same, or must everywhere differ by the same amount. If they are the same, the amplitude of vibration for each particle will be double of that due to either undulation separately. If they are opposite—in other words, if one undulation be half a wave-length in advance of the other—the motions which they would separately produce in any particle are equal and opposite, and the particle will accordingly remain at rest. Two sounds will thus, by their conjoint action, produce silence.

In order that the extinction of sound may be complete, the rarefied portions of each set of waves must be the *exact* counterparts of the condensed portions of the other set, a condition which can only be approximately attained in practice.

The following experiment, due to M. Desains, affords a very direct illustration of the principle of interference. The bottom of a wooden box is pierced with an opening, in which a powerful whistle fits. The top of the box has two larger openings symmetrically placed with respect to the lower one. The inside of the box is lined with felt, to prevent the vibrations from being communicated to the box, and to weaken internal reflection. When the whistle is sounded, if a membrane, with sand strewn on it, is held in various positions in the vertical plane which bisects, at right angles, the line joining the two openings, the sand will be agitated, and will arrange itself in nodal lines. But if it is carried out of this plane, positions will be found, at equal distances on both sides of it, at which the agitation is scarcely perceptible. If, when the membrane is in one of these positions, we close one of the two openings, the sand is again agitated, clearly showing that the previous absence of agitation was due to the interference of the undulations proceeding from the two orifices.

In this experiment the proof is presented to the eye. In the following experiment, which is due to M. Lissajous, it is presented to the ear. A circular plate, supported like the plate in Fig. 594, is made to vibrate in sectors separated by radial nodes. The number of sectors will always be even, and adjacent sectors will vibrate

in opposite directions. Let a disk of card-board of the same size be divided into the same number of sectors, and let alternate sectors be cut away, leaving only enough near the centre to hold the remaining sectors together. If the card be now held just over the vibrating disk, in such a manner that the sectors of the one are exactly over those of the other, a great increase of loudness will be observed, consequent on the suppression of the sound from alternate sectors; but if the card-board disk be turned through the width of half a sector, the effect no longer occurs. If the card is made to rotate rapidly in a continuous manner, the alterations of loudness will form a series of beats.

It is for a similar reason that, when a large bell is vibrating, a person in its centre hears the sound as only moderately loud, while within a short distance of some portions of the edge the loudness is intolerable.

887. **Interference of Direct and Reflected Waves.**[1] **Nodes and Antinodes.**—Interference may also occur between undulations travelling in opposite directions; for example, between a direct and a reflected system. When waves proceeding along a tube meet a rigid obstacle, forming a cross section of the tube, they are reflected directly back again, the motion of any particle close to the obstacle being compounded of that due to the direct wave, and an equal and opposite motion due to the reflected wave. The reflected waves are in fact the images (with reference to the obstacle regarded as a plane mirror) of the waves which would exist in the prolongation of the tube if the obstacle were withdrawn. At the distance of half a wave-length from the obstacle the motions due to the direct and reflected waves will accordingly be equal and opposite, so that the particles situated at this distance will be permanently at rest; and the same is true at the distance of any number of half wave-lengths from the obstacle. The air in the tube will thus be divided into a number of vibrating segments separated by nodal planes or cross sections of no vibration arranged at distances of half a wave-length apart. One of these nodes is at the obstacle itself. At the centres of the vibrating segments—that is to say, at the distance of a quarter wave-length *plus* any number of half wave-lengths from the obstacle or from any node —the velocities due to the direct and reflected waves will be equal and in the same direction, and the amplitude of vibration will accordingly be double of that due to the direct wave alone. These

[1] See note C, page 895.

are the sections of greatest disturbance as regards change of place. We shall call them *antinodes.* On the other hand, it is to be remembered that motion *with* the direct wave is motion *against* the reflected waves, and *vice versa,* so that (§ 869) at points where the velocities due to both have the same absolute direction they correspond to condensation in the case of one of these undulations, and to rarefaction in the case of the other. Accordingly, these sections of maximum movement are the places of no change of density; and on the other hand, the nodes are the places where the changes of density are greatest. If the reflected undulation is feebler than the direct one, as will be the case, for example, if the obstacle is only imperfectly rigid, the destruction of motion at the nodes and of change of density at the antinodes will not be complete; the former will merely be places of minimum motion, and the latter of minimum change of density.

Direct experiments in verification of these principles, a wall being the reflecting body, were conducted by Savart, and also by Seebeck, the latter of whom employed a testing apparatus called the acoustic pendulum. It consists essentially of a small membrane stretched in a frame, from the top of which hangs a very light pendulum, with its bob resting against the centre of the membrane. In the middle portions of the vibrating segments the membrane, moving with the air on its two faces, throws back the pendulum, while it remains nearly free from vibration at the nodes.

Regnault made extensive use of the acoustic pendulum in his experiments on the velocity of sound. The pendulum, when thrown back by the membrane, completed an electric circuit, and thus effected a record of the instant when the sound arrived.

888. Beats Produced by Interference.—When two notes which are not quite in unison are sounded together, a peculiar palpitating effect is produced;—we hear a series of bursts of sound, with intervals of comparative silence between them. The bursts of sound are called *beats,* and the notes are said to *beat* together. If we have the power of tuning one of the notes, we shall find that as they are brought more nearly into unison, the beats become slower, and that, as the departure from unison is increased, the beats become more rapid, till they degenerate first into a rattle, and then into a discord. The effect is most striking with deep notes.

These beats are completely explained by the principle of interference. As the wave-lengths of the two notes are slightly different,

while the velocity of propagation is the same, the two systems of waves will, in some portions of their course, agree in phase, and thus strengthen each other; while in other parts they will be opposite in phase, and will thus destroy each other. Let one of the notes, for example, have 100 vibrations per second, and the other 101. Then, if we start from an instant when the maxima of condensation from the two sources reach the ear together, the next such conjunction will occur exactly a second later. During the interval the maxima of one system have been gradually falling behind those of the other, till, at the end of the second, the loss has amounted to one wave-length. At the middle of the second it will have amounted to half a wave-length, and the two sounds will destroy each other. We shall thus have one beat and one extinction in each second, as a consequence of the fact that the higher note has made one vibration more than the lower. In general, the frequency of beats is the difference of the frequencies of vibration of the beating notes.

NOTE A. § 869.

That the particles which are moving forward are in a state of compression, may be shown in the following way:—Consider an imaginary cross section travelling forward through the tube with the same velocity as the undulation. Call this velocity v, and the velocity of any particle of air u. Also let the density of any particle be denoted by ρ. Then u and ρ remain constant for the imaginary moving section, and the mass of air which it traverses in its motion per unit time is $(v-u)\,\rho$. As there is no permanent transfer of air in either direction through the tube, the mass thus traversed must be the same as if the air were at rest at its natural density. Hence the value of $(v-u)\,\rho$ is the same for all cross sections; whence it follows, that where u is greatest ρ must be greatest, and where u is negative ρ is less than the natural density.

If ρ_0 denote the natural density, we have $(v-u)\,\rho=v\,\rho_0$, whence $\frac{u}{v}=\frac{\rho-\rho_0}{\rho}$; that is to say, *the ratio of the velocity of a particle to the velocity of the undulation is equal to the condensation existing at the particle.* If u is negative—that is to say, if the velocity be retrograde —its ratio to v is a measure of the rarefaction.

From this principle we may easily derive a formula for the velocity of sound, bearing in mind that u is always very small in comparison with v.

For, consider a thin lamina of air whose thickness is δx, and let δu, $\delta\rho$, and δp be the excesses of the velocity, density, and pressure on the second side of the lamina above those on the first at the same moment. The above equation, $(v-u)\,\rho=v\,\rho_0$, gives $(v-u)\,\delta\rho-\rho\,\delta u=0$, whence $\frac{\delta u}{\delta\rho}=\frac{v-u}{\rho}$, or, since u may be neglected in comparison with v,

$$\frac{\delta u}{\delta\rho}=\frac{v}{\rho}.$$

The time which the moving section occupies in traversing the lamina is $\frac{\delta x}{v}$, and in this time the velocity of the lamina changes by the amount $-\delta u$, since the velocity on the

second side of the lamina is $u+\delta u$ at the beginning and u at the end of the time. The force producing this change of velocity (if the section of the tube be unity) is $-\delta p$, or $-1{\cdot}41\,\frac{p}{\rho}\,\delta\rho$, and must be equal to the quotient of change of momentum by time, that is to $-\rho\,\delta x\,.\,\delta u \div \frac{\delta x}{v}$, or to $-\rho\, v\,\delta u$. Hence $\frac{\delta u}{\delta \rho}=1{\cdot}41\,\frac{p}{\rho^2 v}$. Equating this to the other expression for $\frac{\delta u}{\delta \rho}$, we have

$$\frac{v}{\rho}=1{\cdot}41\,\frac{p}{\rho^2 v}\qquad,\qquad v^2=1{\cdot}41\,\frac{p}{\rho}.$$

This investigation is due to Professor Rankine, *Phil. Trans.* 1869.

Note B. § 874.

The following is the usual investigation of the velocity of transmission of sound through a uniform tube filled with air, friction being neglected: Let x denote the original distance of a particle of air from the section of the tube at which the sound originates, and $x+y$ its distance at time t, so that y is the displacement of the particle from the position of equilibrium. Then a particle which was originally at distance $x+\delta x$ will at time t be at the distance $x+\delta x+y+\delta y$; and the thickness of the intervening lamina, which was originally δx, is now $\delta x+\delta y$. Its compression is therefore $-\frac{\delta y}{\delta x}$ or ultimately $-\frac{dy}{dx}$, and if P denote the original pressure, the increase of pressure is $-1{\cdot}41\,\mathrm{P}\,\frac{dy}{dx}$. The excess of pressure behind a lamina δx above the pressure in front is $\frac{d}{dx}\left(1{\cdot}41\,\mathrm{P}\,\frac{dy}{dx}\right)\delta x$, or $1{\cdot}41\,\mathrm{P}\,\frac{d^2 y}{dx^2}\,\delta x$; and if D denote the original density of the air, the acceleration of the lamina will be the quotient of this expression by $\mathrm{D}\,.\,\delta x$. But this acceleration is $\frac{d^2 y}{dt^2}$. Hence we have the equation

$$\frac{d^2 y}{dt^2}=1{\cdot}41\,\frac{\mathrm{P}}{\mathrm{D}}\,\frac{d^2 y}{dx^2};$$

the integral of which is

$$y=\mathrm{F}\,(x-vt)+f\,(x+vt);$$

where v denotes $\sqrt{1{\cdot}41\,\frac{\mathrm{P}}{\mathrm{D}}}$, and F, f denote any functions whatever.

The term $\mathrm{F}\,(x-vt)$ represents a wave, of the form $y=\mathrm{F}\,(x)$, travelling forwards with velocity v; for it has the same value for $t_1+\delta t$ and $x_1+v\,.\,\delta t$ as for t_1 and x_1. The term $f\,(x+vt)$ represents a wave, of the form $y=f\,(x)$, travelling backwards with the same velocity.

In order to adapt this investigation, as well as that given in Note A, to the propagation of longitudinal vibrations through any elastic material, whether solid, liquid, or gaseous, we have merely to introduce E in the place of 1·41 P, E denoting the coefficient of elasticity of the substance, as defined by the condition that a compression $\frac{dy}{dx}$ is produced by a force (per unit area) of $\mathrm{E}\,\frac{dy}{dx}$.

Note C. § 887.

The following is the regular mathematical investigation of the interference of direct and reflected waves of the simplest type, in a uniform tube.

Using x, y, and t in the same sense as in Note B, and measuring x from the reflecting surface to meet the incident waves, we have, for the incident waves,

$$y_1 = a \sin \frac{x + v t}{\lambda} 2\pi, \tag{1}$$

a denoting the amplitude, and λ the wave-length. For the reflected waves, we have

$$y_2 = a \sin \frac{x - v t}{\lambda} 2\pi, \tag{2}$$

since this equation represents waves equal and opposite to the former, and satisfies the condition that at the reflecting surface (where x is zero) the total disturbance $y_1 + y_2$ is zero. Putting y for $y_1 + y_2$, we have, by adding the above equations and employing a well-known formula of trigonometry,

$$y = 2a \sin \frac{x}{\lambda} 2\pi \,.\, \cos \frac{v t}{\lambda} 2\pi. \tag{3}$$

The extension (or compression if negative) is $\frac{dy}{dx}$, and we have

$$\frac{dy}{dx} = \frac{2\pi}{\lambda} 2a \cos \frac{x}{\lambda} 2\pi \,.\, \cos \frac{vt}{\lambda} 2\pi. \tag{4}$$

The factor $\sin \frac{x}{\lambda} 2\pi$ vanishes at the points for which x is either zero or a multiple of $\frac{1}{2}\lambda$, and attains its greatest values (in arithmetical sense) at those for which x is $\frac{1}{4}\lambda$, or $\frac{1}{4}\lambda$ *plus* a multiple of $\frac{1}{2}\lambda$. On the other hand, the factor $\cos \frac{x}{\lambda} 2\pi$ vanishes at the latter points, and attains its greatest values at the former. The points for which $\sin \frac{x}{\lambda} 2\pi$ vanishes are the nodes, since at these points y is constantly zero; and the points for which $\cos \frac{x}{\lambda} 2\pi$ vanishes are the antinodes, since at these the extension or compression is constantly zero.

The motion represented by equation (3) is the simplest type of *stationary undulation.*

CHAPTER LXIII.

NUMERICAL EVALUATION OF SOUND.

889. Qualities of Musical Sound.—Musical tones differ one from another in respect of three qualities;—loudness, pitch, and character.

Loudness.—The loudness of a sound considered subjectively is the intensity of the sensation with which it affects the organs of hearing. Regarded objectively, it depends, in the case of sounds of the same pitch and character, upon the energy of the aerial vibrations in the neighbourhood of the ear, and is proportional to the square of the amplitude.

Our auditory apparatus is, however, so constructed as to be more susceptible of impression by sounds of high than of low pitch. A bass note must have much greater energy of vibration than a treble note, in order to strike the ear as equally loud. The intensity of sonorous vibration at a point in the air is therefore not an absolute measure of the intensity of the sensation which will be received by an ear placed at the point.

The word loud is also frequently applied to a source of sound, as when we say a loud voice, the reference being to the loudness as heard at a given distance from the source. The diminution of loudness with increase of distance according to the law of inverse squares is essentially connected with the proportionality of loudness to square of amplitude.

Pitch.—Pitch is the quality in respect of which an acute sound differs from a grave one; for example, a treble note from a bass note. All persons are capable of appreciating differences of pitch to some extent, and the power of forming accurate judgments of pitch constitutes what is called a *musical ear.*

Physically, pitch depends solely on *frequency of vibration*, that is to say, on the number of vibrations executed per unit time. In

ordinary circumstances this frequency is the same for the source of sound, the medium of transmission, and the drum of the ear of the person hearing; and in general the transmission of vibrations from one body or medium to another produces no change in their frequency. The *second* is universally employed as the unit of time in treating of sonorous vibrations; so that *frequency* means *number of vibrations per second.* Increase of frequency corresponds to elevation of pitch.

Period and *frequency* are reciprocals. For example, if the period of each vibration is $\frac{1}{100}$ of a second, the number of vibrations per second is 100. Period therefore is an absolute measure of pitch, and the longer the period the lower is the pitch.

The wave-length of a note in any medium is the distance which sound travels in that medium during the period corresponding to the note. Hence wave-length may be taken as a measure of pitch, provided the medium be given; but, in passing from one medium to another, wave-length varies directly as the velocity of sound. The wave-length of a given note in air depends upon the temperature of the air, and is shortened in transmission from the heated air of a concert-room to the colder air outside, while the pitch undergoes no change.

If we compare a series of notes rising one above another by what musicians regard as equal differences of pitch, their frequencies will not be equidifferent, but will form an increasing geometrical progression, and their periods (and wave-lengths in a given medium) will form a decreasing geometrical progression.

Character.—Musical sounds may, however, be alike as regards pitch and loudness, and may yet be easily distinguishable. We speak of the *quality* of a singer's voice, and the *tone* of a musical instrument; and we characterize the one or the other as rich, sweet, or mellow; on the one hand: or as poor, harsh, nasal, &c., on the other. These epithets are descriptive of what musicians call *timbre*—a French word literally signifying *stamp*. German writers on acoustics denote the same quality by a term signifying *sound-tint*. It might equally well be called *sound-flavour*. We adopt *character* as the best English designation.

Physically considered, as wave-length and wave-amplitude fall under the two previous heads, *character* must depend upon the only remaining point in which aerial waves can differ—namely their *form*, meaning by this term the law according to which the velo-

cities and densities change from point to point of a wave. This subject will be more fully treated in Chapter lxv. Every musical sound is more or less mingled with non-musical noises, such as puffing, scraping, twanging, hissing, rattling, &c. These are not comprehended under *timbre* or *character* in the usage of the best writers on acoustics. The gradations of loudness which characterize the commencement, progress, and cessation of a note, and upon which musical effect often greatly depends, are likewise excluded from this designation. In distinguishing the sounds of different musical instruments, we are often guided as much by these gradations and extraneous accompaniments as by the character of the musical tones themselves.

890. Musical Intervals.—When two notes are heard, either simultaneously or in succession, the ear experiences an impression of a special kind, involving a perception of the relation existing between them as regards difference of pitch. This impression is often recognized as identical where absolute pitch is very different, and we express this identity of impression by saying that the *musical interval* is the same.

Each musical interval, thus recognized by the ear as constituting a particular relation between two notes, is found to correspond to a particular *ratio* between their frequencies of vibration. Thus the *octave*, which of all intervals is that which is most easily recognized by the ear, is the relation between two notes whose *frequencies* are as 1 to 2, the upper note making twice as many vibrations as the lower in any given time.

It is the musician's business so to combine sounds as to awaken emotions of the peculiar kind which are associated with works of art. In attaining this end he employs various resources, but musical intervals occupy the foremost place. It is upon the judicious employment of these that successful composition mainly depends.

891. Gamut.—The *gamut* or *diatonic scale* is a series of eight notes having certain definite relations to one another as regards frequency of vibration. The first and last of the eight are at an interval of an octave from each other, and are called by the same name; and by taking in like manner the octaves of the other notes of the series, we obtain a repetition of the gamut both upwards and downwards, which may be continued over as many octaves as we please.

The notes of the gamut are usually called by the names

Do	Re	Mi	Fa	Sol	La	Si	Do_2

and their vibration-frequencies are proportional to the numbers

1	$\frac{9}{8}$	$\frac{5}{4}$	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{15}{8}$	2

or, clearing fractions, to

24	27	30	32	36	40	45	48

The intervals from Do to each of the others in order are called a *second*, a *major third*, a *fourth*, a *fifth*, a *sixth*, a *seventh*, and an *octave* respectively. The interval from La to Do_2 is called a *minor third*, and is evidently represented by the ratio $\frac{6}{5}$.

The interval from Do to Re, from Fa to Sol, or from La to Si, is represented by the ratio $\frac{9}{8}$, and is called a *major tone*. The interval from Re to Mi, or from Sol to La, is represented by the ratio $\frac{10}{9}$, and is called a *minor tone*. The interval from Mi to Fa, or from Si to Do_2, is represented by the ratio $\frac{16}{15}$, and is called a *limma*. As the square of $\frac{16}{15}$ is a little greater than $\frac{9}{8}$, a limma is rather more than half a major tone.

The intervals between the successive notes of the gamut are accordingly represented by the following ratios[1]:—

Do		Re		Mi		Fa		Sol		La		Si		Do_2
	$\frac{9}{8}$		$\frac{10}{9}$		$\frac{16}{15}$		$\frac{9}{8}$		$\frac{10}{9}$		$\frac{9}{8}$		$\frac{16}{15}$	

Do (with all its octaves) is called the *key-note* of the piece of music, and may have any pitch whatever. In order to obtain perfect harmony, the above ratios should be accurately maintained whatever the key-note may be.

892. Tempered Gamut.—A great variety of keys are employed in music, and it is a practical impossibility, at all events in the case of instruments like the piano and organ, which have only a definite set of notes, to maintain these ratios strictly for the whole range of possible key-notes. Compromise of some kind becomes necessary, and different systems of compromise are called different *temperaments* or different *modes of temperament*. The temperament which is most in favour in the present day is the simplest possible, and is called *equal temperament*, because it favours no key above another, but makes the tempered gamut exactly the same for all. It ignores the

[1] The logarithmic differences, which are accurately proportional to the intervals, are approximately as under, omitting superfluous zeros.

Do		Re		Mi		Fa		Sol		La		Si		Do
	51		46		28		51		46		51		28	

difference between major and minor tones, and makes the limma exactly half of either. The interval from Do to Do_2 is thus divided into 5 tones and 2 semitones, a tone being $\frac{1}{6}$ of an octave, and a semitone $\frac{1}{12}$ of an octave. The ratio of frequencies corresponding to a tone will therefore be the sixth root of 2, and for a semitone it will be the 12th root of 2.

The difference between the natural and the tempered gamut for the key of C is shown by the following table, which gives the number of complete vibrations per second for each note of the middle octave of an ordinary piano:—

	Tempered Gamut.	Natural Gamut.		Tempered Gamut.	Natural Gamut.
C . .	258·7	258·7	G . .	387·6	388·0
D . .	290·3	291·0	A . .	435·0	431·1
E . .	325·9	323·4	B . .	488·2	485·0
F . .	345·3	344·9	C . .	517·3	517·3

The absolute pitch here adopted is that of the Paris Conservatoire, and is fixed by the rule that A (the middle A of a piano, or the A string of a violin) is to have 435 complete vibrations per second in the tempered gamut. This is rather lower than the concert-pitch which has prevailed in this country in recent years, but is probably not so low as that which prevailed in the time of Handel. It will be noted that the number of vibrations corresponding to C is approximately equal to a power of 2 (256 or 512). Any power of 2 accordingly expresses (to the same degree of approximation) the number of vibrations corresponding to one of the octaves of C.

The Stuttgard congress (1834) recommended 528 vibrations per second for C, and the C tuning-forks sold under the sanction of the Society of Arts are guaranteed to have this pitch. By multiplying the numbers 24, 27 . . . 48, in § 891, by 11, we shall obtain the frequencies of vibration for the natural gamut in C corresponding to this standard. What is generally called *concert-pitch* gives C about 538. The C of the Italian Opera is 546. Handel's C is said to have been $499\frac{1}{5}$.

893. Limits of Pitch employed in Music.—The deepest note regularly employed in music is the C of 32 vibrations per second which is emitted by the longest pipe (the 16-foot pipe) of most organs. Its wave-length in air at a temperature at which the velocity of sound is 1120 feet per second, is $\frac{1120}{32} = 35$ feet. The highest note employed seldom exceeds A, the third octave of the A above defined. Its number of vibrations per second is $435 \times 2^3 = 3480$, and

its wave-length in air is about 4 inches. Above this limit it is difficult to appreciate pitch, but notes of at least ten times this number of vibrations are audible.

The average compass of the human voice is about two octaves. The deep F of a bass-singer has 87, and the upper G of the treble 775 vibrations per second. Voices which exceed either of these limits are regarded as deep or high.

894. **Minor Scale and Pythagorean Scale.**—The difference between a major and minor tone is expressed by the ratio $\frac{81}{80}$, and is called a *comma*. The difference between a minor tone and a limma is expressed by the ratio $\frac{25}{24}$, and is the smallest value that can be assigned to the somewhat indefinite interval denoted by the same *semitone*, the greatest value being the limma itself ($\frac{16}{15}$). The signs ♯ and ♭ (sharp and flat) appended to a note indicate that it is to be raised or lowered by a semitone. The major scale or gamut, as above given, is modified in the following way to obtain the minor scale:—

Do		Re		Mi♭		Fa		Sol		La♭		Si♭		Do$_2$
	$\frac{9}{8}$		$\frac{16}{15}$		$\frac{10}{9}$		$\frac{9}{8}$		$\frac{16}{15}$		$\frac{9}{8}$		$\frac{10}{9}$	

the numbers in the second line being the ratios which represent the intervals between the successive notes.

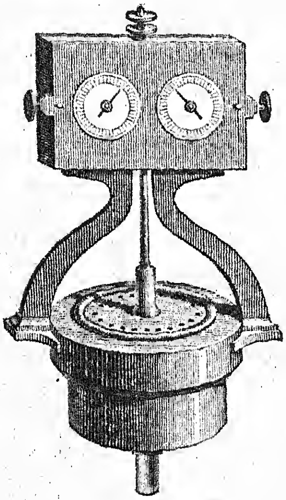
It is worthy of note that Pythagoras, who was the first to attempt the numerical evaluation of musical intervals, laid down a scheme of values slightly different from that which is now generally adopted. According to him, the intervals between the successive notes of the major scale are as follows:—

Do		Re		Mi		Fa		Sol		La		Si		Do
	$\frac{9}{8}$		$\frac{9}{8}$		$\frac{256}{243}$		$\frac{9}{8}$		$\frac{9}{8}$		$\frac{9}{8}$		$\frac{256}{243}$	

This scheme agrees exactly with the common system as regards the values of the fourth, fifth, and octave, and makes the values of the major third, the sixth, and the seventh each greater by a comma, while the small interval from *mi* to *fa*, or from *si* to *do*, is diminished by a comma. In the ordinary system, the prime numbers which enter the ratios are 2, 3, and 5; in the Pythagorean system they are only 2 and 3; hence the interval between any two notes of the Pythagorean scale can be expressed as the sum or difference of a certain number of octaves and fifths. In tuning a violin by making the intervals between the strings true fifths, the Pythagorean scheme is virtually employed.

895. Methods of Counting Vibrations. Siren.—The instrument which is chiefly employed for counting the number of vibrations corresponding to a given note, is called the *siren*, and was devised by Cagniard de Latour. It is represented in Figs. 614, 615, the former being a front, and the latter a back view.

There is a small wind-chest, nearly cylindrical, having its top pierced with fifteen holes, disposed at equal distances round the circumference of a circle. Just over this, and nearly touching it, is a movable circular plate, pierced with the same number of holes



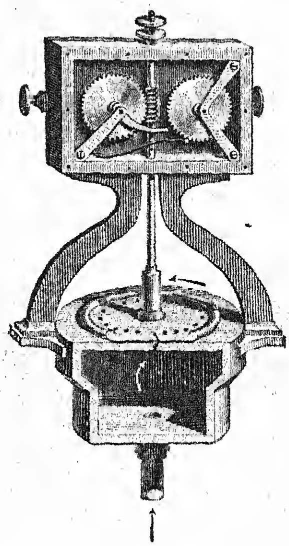


Fig. 614. Siren. Fig. 615.

similarly arranged, and so mounted that it can rotate very freely about its centre, carrying with it the vertical axis to which it is attached. This rotation is effected by the action of the wind, which enters the wind-chest from below, and escapes through the holes. The form of the holes is shown by the section in Fig. 615. They do not pass perpendicularly through the plates, but slope contrary ways, so that the air when forced through the holes in the lower plate impinges *upon one side* of the holes in the upper plate, and thus blows it round in a definite direction. The instrument is driven by means of the bellows shown in Fig. 625 (§ 910). As the rotation of one plate upon the other causes the holes to be alternately opened and closed, the wind escapes in successive puffs, whose frequency

depends upon the rate of rotation. Hence a note is emitted which rises in pitch as the rotation becomes more rapid.

The siren will sound under water, if water is forced through it instead of air; and it was from this circumstance that it derived its name.

In each revolution, the fifteen holes in the upper plate come opposite to those in the lower plate 15 times, and allow the compressed air in the wind-chest to escape; while in the intervening positions its escape is almost entirely prevented. Each revolution thus gives rise to 15 vibrations; and in order to know the number of vibrations corresponding to the note emitted, it is only necessary to have a means of counting the revolutions.

This is furnished by a counter, which is represented in Fig. 615. The revolving axis carries an endless screw, driving a wheel of 100 teeth, whose axis carries a hand traversing a dial marked with 100 divisions. Each revolution of the perforated plate causes this hand to advance one division. A second toothed-wheel is driven intermittently by the first, advancing suddenly one tooth whenever the hand belonging to the first wheel passes the zero of its scale. This second wheel also carries a hand traversing a second dial; and at each of the sudden movements just described this hand advances one division. Each division accordingly indicates 100 revolutions of the perforated plate, or 1500 vibrations. By pushing in one of the two buttons which are shown, one on each side of the box containing the toothed-wheels, we can instantaneously connect or disconnect the endless screw and the first toothed-wheel.

In order to determine the number of vibrations corresponding to any given sound which we have the power of maintaining steadily, we fix the siren on the bellows, the screw and wheel being disconnected, and drive the siren until the note which it emits is judged to be in unison with the given note. We then, either by regulating the pressure of the wind, or by employing the finger to press with more or less friction against the revolving axis, contrive to keep the note of the siren constant for a measured interval of time, which we observe by a watch. At the commencement of the interval we suddenly connect the screw and toothed-wheel, and at its termination we suddenly disconnect them, having taken care to keep the siren in unison with the given sound during the interval. As the hands do not advance on the dials when the screw is out of connection with the wheels, the readings before and after the measured interval of

time can be taken at leisure. Each reading consists of four figures, indicating the number of revolutions from the zero position, units and tens being read off on the first dial, and hundreds and thousands on the second. The difference of the two readings is the number of revolutions made in the measured interval, and when multiplied by 15 gives the number of vibrations in the interval, whence the number of vibrations per second is computed by division.

896. Graphic Method.—In the hands of a skilful operator, with a good musical ear, the siren is capable of yielding very accurate determinations, especially if, by adding or subtracting the number of beats,

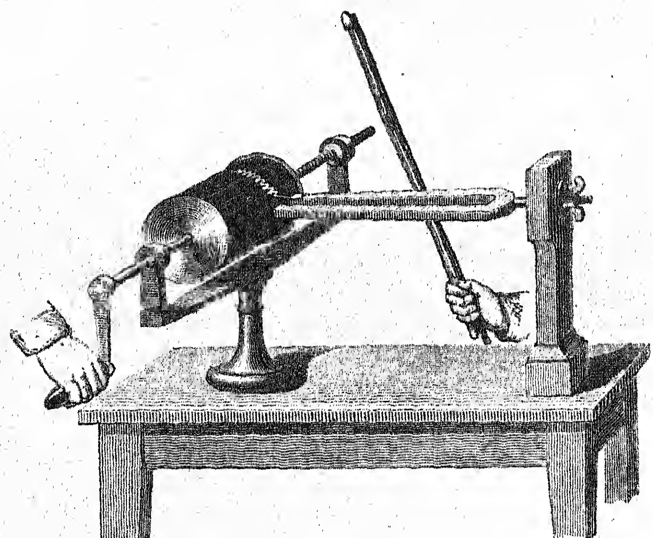


Fig. 616.—Vibroscope.

correction be made for any slight difference of pitch between the siren and the note under investigation.

The vibrations of a tuning-fork can be counted, without the aid of the siren, by a graphical method, which does not call for any exercise of musical judgment, but simply involves the performance of a mechanical operation.

The tuning-fork is fixed in a horizontal position, as shown in Fig. 616, and has a light style, which may be of brass wire, quill, or bristle, attached to one of its prongs, by wax or otherwise. To receive the trace, a piece of smoked paper is gummed round a cylinder, which can be turned by a handle, a screw cut on the axis

causing it at the same time to travel endwise. The cylinder is placed so that the style barely touches the blackened surface. The fork is then made to vibrate by bowing it, and the cylinder is turned. The result is a wavy line traced on the blackened surface, and the number of wave-forms (each including a pair of bends in opposite directions) is the number of vibrations. If the experiment lasts for a measured interval of time, we have only to count these wave-forms, and divide by the number of seconds, in order to obtain the number of vibrations per second for the note of the tuning-fork. By plunging the paper in ether, the trace will be fixed, so that the paper may be laid aside, and the vibrations counted at leisure. The apparatus is called the *vibroscope,* and was invented by Duhamel.

M. Léon Scott has invented an instrument called the *phonautograph,* which is adapted to the graphical representation of sounds in

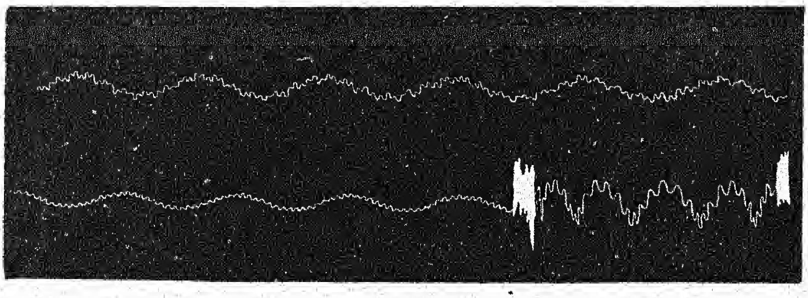


Fig. 617.—Traces by Phonautograph.

general. The style, which is very light, is attached to a membrane stretched across the smaller end of what may be called a large ear-trumpet. The membrane is agitated by the aerial waves proceeding from any source of sound, and the style leaves a record of these agitations on a blackened cylinder, as in Duhamel's apparatus. Fig. 617 represents the traces thus obtained from the sound of a tuning-fork in three different modes of vibration.

897. **Tonometer.**—When we have determined the frequency of vibration for a particular tuning-fork, that of another fork, nearly in unison with it, can be deduced by making the two forks vibrate simultaneously, and counting the beats which they produce.

Scheibler's *tonometer,* which is constructed by Koenig of Paris, consists of a set of 65 tuning-forks, such that any two consecutive forks make 4 beats per second, and consequently differ in pitch by

4 vibrations per second. The lowest of the series makes 256 vibrations, and the highest 512, thus completing an octave. Any note within this range can have its vibration-frequency at once determined, with great accuracy, by making it sound simultaneously with the fork next above or below it, and counting beats.

With the aid of this instrument, a piano can be tuned with certainty to any desired system of temperament, by first tuning the notes which come within the compass of the tonometer, and then proceeding by octaves.

In the ordinary methods of tuning pianos and organs, temperament is to a great extent a matter of chance; and a tuner cannot attain the same temperament in two successive attempts.

898. Pitch modified by Relative Motion.—We have stated in § 889 that, in ordinary circumstances, the frequency of vibration in the source of sound, is the same as in the ear of the listener, and in the intervening medium. This identity, however, does not hold if the source of sound and the ear of the listener are approaching or receding from each other. Approach of either to the other produces increased frequency of the pulses on the ear, and consequent elevation of pitch in the sound as heard; while recession has an opposite effect. Let n be the number of vibrations performed in a second by the source of the sound, v the velocity of sound in the medium, and a the relative velocity of approach. Then the number of waves which reach the ear of the listener in a second, will be n *plus* the number of waves which cover a length a, that is (since n waves cover a length v), will be $n+\frac{a}{v}n$ or $\frac{v+a}{v}n$.

The following investigation is more rigorous. Let the source make n vibrations per second. Let the observer move towards the source with velocity a. Let the source move away from the observer with velocity a'. Let the medium move from the observer towards the source with velocity m, and let the velocity of sound in the medium be v.

Then the velocity of the observer relative to the medium is $a-m$ towards the source, and the velocity of the source relative to the medium is $a'-m$ away from the observer. The velocity of the sound relative to the source will be different in different directions, its greatest amount being $v+a'-m$ towards the observer, and its least being $v-a'+m$ away from the observer. The length of a wave will vary with direction, being $\frac{1}{n}$ of the velocity of the sound

relative to the source. The length of those waves which meet the observer will be $\frac{v+a'-m}{n}$, and the velocity of these waves relative to the observer will be $v+a-m$; hence the number of waves that meet him in a second will be $\frac{v+a-m}{v+a'-m}n$.

Careful observation of the sound of a railway whistle, as an express train dashes past a station, has confirmed the fact that the sound as heard by a person standing at the station is higher while the train is approaching than when it is receding. A speed of about 40 miles an hour will sharpen the note by a semitone in approaching, and flatten it by the same amount in receding, the natural pitch being heard at the instant of passing.[1]

[1] The best observations of this kind were those of Buys Ballot, in which trumpeters, with their instruments previously tuned to unison, were stationed, one on the locomotive, and others at three stations beside the line of railway. Each trumpeter was accompanied by musicians, charged with the duty of estimating the difference of pitch between the note of his trumpet and those of the others, as heard before and after passing.

CHAPTER LXIV.

MODES OF VIBRATION.

899. Longitudinal and Transverse Vibrations of Solids.—Sonorous vibrations are manifestations of elasticity. When the particles of a solid body are displaced from their natural positions relative to one another by the application of external force, they tend to return, in virtue of the elasticity of the body. When the external force is removed, they spring back to their natural position, pass it in virtue of the velocity acquired in the return, and execute isochronous vibrations about it until they gradually come to rest. The isochronism of the vibrations is proved by the constancy of pitch of the sound emitted; and from the isochronism we can infer, by the aid of mathematical reasoning, that the restoring force increases directly as the displacement of the parts of the body from their natural relative position (§ 111).

The same body is, in general, susceptible of many different modes of vibration, which may be excited by applying forces to it in different ways. The most important of these are comprehended under the two heads of *longitudinal* and *transverse* vibrations.

In the former the particles of the body move to and fro in the direction along which the pulses travel, which is always regarded as the longitudinal direction, and the deformations produced consist in alternate compressions and extensions. In the latter the particles move to and fro in directions transverse to that in which the pulses travel, and the deformation consists in bending. To produce longitudinal vibrations, we must apply force in the longitudinal direction. To produce transverse vibration, we must apply force transversely.

900. Transverse Vibrations of Strings.—To the transverse vibrations of strings, instrumental music is indebted for some of its most

precious resources. In the violin, violoncello, &c., the strings are set in vibration by drawing a bow across them. The part of the bow which acts on the strings consists of hairs tightly stretched and rubbed with rosin. The bow adheres to the string, and draws it aside till the reaction becomes too great for the adhesion to overcome. As the bow continues to be drawn on, slipping takes place, and the mere fact of slipping diminishes the adhesion. The string accordingly springs back suddenly through a finite distance. It is then again caught by the bow, and the same action is repeated. In the harp and guitar, the strings are plucked with the finger, and then left to vibrate freely. In the piano the wires are struck with little hammers faced with leather. The pitch of the sound emitted in these various cases depends only on the string itself, and is the same whichever mode of excitation be employed.

901. Laws of the Transverse Vibrations of Strings.—It can be shown by an investigation closely analogous to that which gives the velocity of sound in air, that the velocity with which transverse vibrations travel along a perfectly flexible string is given by the formula

$$v=\sqrt{\frac{t}{m}}; \qquad (1)$$

t denoting the tension of the string, and m the mass of unit length of it. If m be expressed in grammes per centimetre of length, t should be expressed in dynes (§ 87), and the value obtained for v will be in centimetres per second. The sudden disturbance of any point in the string, causes two pulses to start from this point, and run along the string in opposite directions. Each of these, on arriving at the end of the free portion of the string, is reflected from the solid support to which the string is attached, and at the same time undergoes reversal as to side. It runs back, thus reversed, to the other end of the free portion, and there again undergoes reflection and reversal. When it next arrives at the origin of the disturbance it has travelled over just twice the length of the string; and as this is true of both the pulses, they must both arrive at this point together. At the instant of their meeting, things are in the same condition as when the pulses were originated, and the movements just described will again take place. The period of a complete vibration of the string is therefore the time required for a pulse to travel over twice its length; that is,

$$\frac{1}{n}=\frac{2l}{v}=2l\sqrt{\frac{m}{t}};$$

$$\text{or } n=\frac{1}{2l}\sqrt{\frac{t}{m}}; \qquad (2)$$

l denoting the length of the string between its points of attachment, and n the number of vibrations per second.

This formula involves the following laws:—

1. When the length of the vibrating portion of the string is altered, without change of tension, the frequency of vibration varies inversely as the length.

2. If the tension be altered, without change of length in the vibrating portion, the frequency of vibration varies as the square root of the tension.

3. Strings of the same length, stretched with the same forces, have frequencies of vibration which are inversely as the square roots of their masses (or weights).

4. Strings of the same length and density, but of different thicknesses, will vibrate in the same time, if they are stretched with forces proportional to their sectional areas.

All these laws are illustrated (qualitatively, if not quantitatively) by the strings of a violin.

The first is illustrated by the fingering, the pitch being raised as the portion of string between the finger and the bridge is shortened.

The second is illustrated by the mode of tuning, which consists in tightening the string if its pitch is to be raised, or slackening the string if it is to be lowered.

The third law is illustrated by the construction of the bass string, which is wrapped round with metal wire, for the purpose of adding to its mass, and thus attaining slow vibration without undue slackness. The tension of this string is in fact greater than that of the string next it, though the latter vibrates more rapidly in the ratio of 3 to 2.

The fourth law is indirectly illustrated by the sizes of the first three strings. The treble string is the smallest, and is nevertheless stretched with much greater force than any of the others. The third string is the thickest, and is stretched with less force than any of the others. The increased thickness is necessary in order to give sufficient power in spite of the slackness of the string.

902. Experimental Illustration: Sonometer.—For the quantitative illustration of these laws, the instrument called the sonometer, represented in Fig. 618, is commonly employed. It consists essen-

tially of a string or wire stretched over a sounding-box by means of a weight. One end of the string is secured to a fixed point at one end of the sounding-box. The other end passes over a pulley, and carries weights which can be altered at pleasure. Near the two ends of the box are two fixed bridges, over which the cord passes. There is also a movable bridge, which can be employed for altering the length of the vibrating portion.

To verify the law of lengths, the whole length between the fixed bridges is made to vibrate, either by plucking or bowing; the movable bridge is then introduced exactly in the middle, and one of the halves is made to vibrate; the note thus obtained will be found to be the upper octave of the first. The frequency of vibration is therefore doubled. By making two-thirds of the whole length vibrate, a note will be obtained which will be recognized as the fifth of the fundamental note, its vibration-frequency being therefore greater in the ratio $\frac{3}{2}$. To obtain the notes of the gamut, we commence with the string as a whole, and then employ portions of its length represented by the fractions $\frac{8}{9}$, $\frac{4}{5}$, $\frac{3}{4}$, $\frac{2}{3}$, $\frac{3}{5}$, $\frac{8}{15}$, $\frac{1}{2}$.

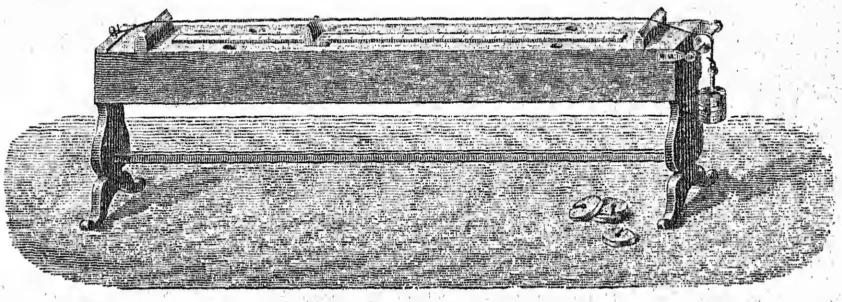


Fig. 618.—Sonometer.

To verify the law independently of all knowledge of musical intervals, a light style may be attached to the cord, and caused to trace its vibrations on the vibroscope. This mode of proof is also more general, inasmuch as it can be applied to ratios which do not correspond to any recognized musical interval.

To verify the law of tensions, we must change the weight. It will be found that, to produce a rise of an octave in pitch, the weight must be increased fourfold.

To verify the third and fourth laws, two strings must be employed, their masses having first been determined by weighing them.

If the strings are thick, and especially if they are thick steel wires, their flexural rigidity has a sensible effect in making the vibrations quicker than they would be if the tension acted alone.

903. **Harmonics.**—Any person of ordinary musical ear may easily, by a little exercise of attention, detect in any note of a piano the presence of its upper octave, and of another note a fifth higher than this; these being the notes which correspond to frequencies of vibration double and triple that of the fundamental note. A highly trained ear can detect the presence of other notes, corresponding to still higher multiples of the fundamental frequency of vibration. Such notes are called *harmonics*.

When the vibration-frequency of one note is an exact multiple of that of another note, the former note is called a harmonic of the latter. The notes of all stringed instruments contain numerous harmonics blended with the fundamental tones. Bells and vibrating plates have higher tones mingled with the fundamental tone; but these higher tones are not harmonics in the sense in which we use the word.

A violin string sometimes fails to yield its fundamental note, and gives the octave or some other harmonic instead. This result can be brought about at pleasure, by lightly touching the string at a properly-selected point in its length, while the bow is applied in the usual way. If touched at the middle point of its length, it gives the octave. If touched at one-third of its length from either end, it gives the fifth above the octave. The law is, that if touched at $\frac{1}{n}$ of its length[1] from either end, it yields the harmonic whose vibration-frequency is n times that of the fundamental tone. The string in these cases divides itself into a number of equal vibrating-segments, as shown in Fig. 619.

The division into segments is often distinctly *visible* when the string of a sonometer is strongly bowed, and its existence can be verified, when less evident, by putting paper riders on different parts of the string. These (as shown in the figure) will be thrown off by the vibrations of the string, unless they are placed accurately at the nodal points, in which case they will retain their seats. If two strings tuned to unison are stretched on the same sonometer, the vibration of the one induces similar vibrations in the other; and the experiment of the riders may be varied, in a very instructive way,

[1] Or at $\frac{m}{n}$ of its length, if m be prime to n.

by bowing one string and placing the riders on the other. This is an instance of a general principle of great importance—that a vibrating body communicates its vibrations to other bodies which are capable of vibrating in unison with it. The propagation of a sound may indeed be regarded as one grand vibration in unison; but, besides the general waves of *propagation,* there are waves of *re-*

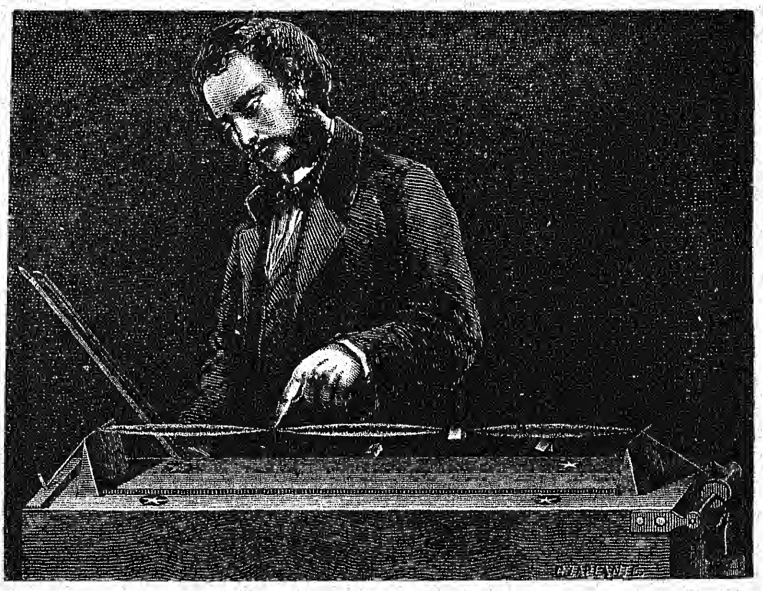


Fig. 619.—Production of a Harmonic.

inforcement, due to the synchronous vibrations of limited portions of the transmitting medium. This is the principle of resonance.

904. Resonance.—By applying to a pendulum originally at rest a series of very feeble impulses, at intervals precisely equal to its natural time of vibration, we shall cause it to swing through an arc of considerable magnitude.

The same principle applies to a body capable of executing vibrations under the influence of its own elasticity. A series of impulses keeping time with its own natural period may set it in powerful vibration, though any one of them singly would have no appreciable effect.

Some bodies, such as strings and confined portions of air, have definite periods in which they can vibrate freely when once started;

and when a note corresponding to one of these periods is sounded in their neighbourhood, they readily take it up and emit a note of the same pitch themselves.

Other bodies, especially thin pieces of dry straight-grained deal, such as are employed for the faces of violins and the sounding-boards of pianos, are capable of vibrating, more or less freely, in any period lying between certain wide limits. They are accordingly set in vibration by all the notes of their respective instruments; and by the large surface with which they act upon the air, they contribute in a very high degree to increase the sonorous effect. All stringed instruments are constructed on this principle; and their quality mainly depends on the greater or less readiness with which they respond to the vibrations of the strings.

All such methods of reinforcing a sound must be included under *resonance;* but the word is often more particularly applied to the reinforcement produced by masses of air.

905. Longitudinal Vibrations of Strings.—Strings or wires may also be made to vibrate *longitudinally*, by rubbing them, in the direction of their length, with a bow or a piece of chamois leather covered with rosin. The sounds thus obtained are of much higher pitch than those produced by transverse vibration.

In the case of the fundamental note, each of the two halves A C, C B (Fig. 620), is alternately extended and compressed, one being

Fig. 620.—Longitudinal Vibration. First Tone.

extended while the other is compressed. At the middle point C there is no extension or compression, but there is greater amplitude of movement than at any other point. The amplitudes diminish in passing from C towards either end, and vanish at the ends, which are therefore nodes. The extensions and compressions, on the other hand, increase as we travel from the middle towards either end, and obtain their greatest values at the ends.

But the string may also divide itself into any number of separately-vibrating segments, just as in the case of transverse vibrations. Fig. 621 represents the motions which occur when there are three such segments, separated by two nodes D, E. The upper portion of the figure is true for one-half of the period of vibration, and the lower portion for the remaining half.

The frequency of vibration, for longitudinal as well as for transverse vibrations, varies inversely as the length of the vibrating string, or segment of string. We shall return to this subject in § 916.

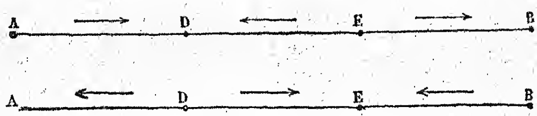


Fig. 621.—Longitudinal Vibration. Third Tone.

906. Stringed Instruments.—Only the transversal vibrations of strings are employed in music. In the violin and violoncello there are four strings, each being tuned a fifth above the next below it; and intermediate notes are obtained by fingering, the portion of string between the finger and the bridge being the only part that is free to vibrate. The bridge and sounding-post serve to transmit the vibrations of the strings to the body of the instrument. In the piano there is also a bridge, which is attached to the sounding-board, and communicates to it the vibrations of the wires.

907. Transversal Vibrations of Rigid Bodies: Rods, Plates, Bells.—We shall not enter into detail respecting the laws of the transverse vibrations of rigid bodies. The relations of their overtones to their fundamental tones are usually of an extremely complex character, and this fact is closely connected with the unmusical or only semi-musical character of the sounds emitted.

When one face of the body is horizontal, the division into separate vibrating segments can be rendered visible by a method devised by Chladni, namely, by strewing sand on this face. During the vibration, the sand, as it is tossed about, works its way to certain definite lines, where it comes nearly to rest. These nodal lines must be regarded as the intersections of internal nodal surfaces with the surface on which the sand is strewed, each nodal surface being the boundary between parts of the body which have opposite motions.

The figures composed by these nodal lines are often very beautiful, and quite startling in the suddenness of their production. Chladni and Savart published the forms of a great number. A complete theoretical explanation of them would probably transcend the powers of the greatest mathematicians.

Bells and bell-glasses vibrate in segments, which are never less than four in number, and are separated by nodal lines meeting in the middle of the crown. They are well shown by putting water in a

bell-glass, and bowing its edge. The surface of the water will immediately be covered with ripples, one set of ripples proceeding from each of the vibrating segments. The division into any possible number of segments may be effected by pressing the glass with the fingers in the places where a pair of consecutive nodes ought to be formed, while the bow is applied to the middle of one of the segments. The greater the number of segments the higher will be the note emitted.

908. Tuning-fork.—Steel rods, on account of their comparative freedom from change, are well suited for standards of pitch. The tuning-fork, which is especially used for this purpose, consists essentially of a steel rod bent double, and attached to a handle of the same material at its centre. Besides the fundamental tone, it is capable of yielding two or three overtones, which are very much higher in pitch; but these are never used for musical purposes. If the fork is held by the handle while vibrating, its motion continues for a long time, but the sound emitted is too faint to be heard except by holding the ear near it. When the handle is pressed against a table, the latter acts as a sounding-board, and communicates the vibrations to the air, but it also causes the fork to come much more speedily to rest. For the purposes of the lecture-room the fork is often mounted on a sounding-box (Fig. 622), which should be separated from the table by two pieces of india-rubber tubing. The box can then vibrate freely in unison with the fork, and the sound is both loud and lasting. The vibrations are usually excited either by bowing the fork or by drawing a piece of wood between its prongs.

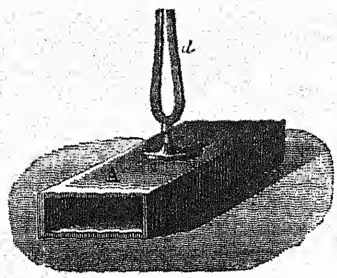


Fig. 622.—Fork on Sounding-box.

The pitch of a tuning-fork varies slightly with temperature, becoming lower as the temperature rises. This effect is due in some trifling degree to expansion, but much more to the diminution of elastic force.

909. Law of Linear Dimensions.—The following law is of very wide application, being applicable alike to solid, liquid, and gaseous bodies: —*When too bodies differing in size, but in other respects similar and similarly circumstanced, vibrate in the same mode, their vibration-periods are directly as their linear dimensions.* Their vibra-

tion-frequencies are consequently in the inverse ratio of their linear dimensions.

In applying the law to the transverse vibrations of strings, it is to be understood that the stretching force per unit of sectional area is constant. In this case the velocity of a pulse (§ 901) is constant, and the period of vibration, being the time required for a pulse to travel over twice the length of the string, is therefore directly as the length.

910. Organ-pipes.—In organs, and wind-instruments generally, the sonorous body is a column of air confined in a tube. To set this air

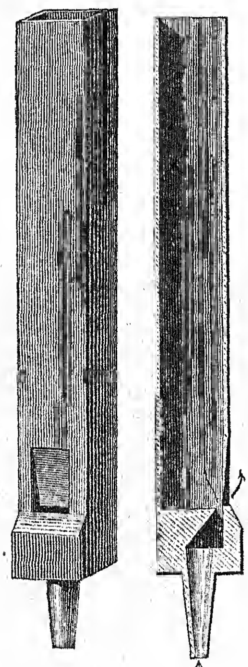


Fig. 623.—Block Pipe.

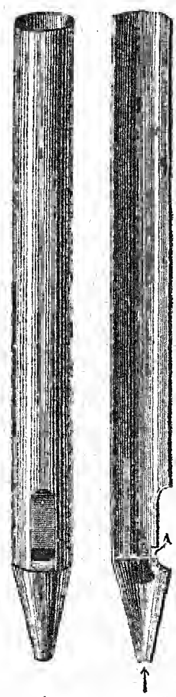


Fig. 624.—Flue Pipe.

in vibration some kind of mouth-piece must be employed. That which is most extensively used in organs is called the *flute mouth-piece*,[1] and is represented, in conjunction with the pipe to which it is attached, in Figs. 623, 624. It closely resembles the mouth-piece of

[1] This is not the trade name. English organ-builders have no generic name for this mouth-piece.

an ordinary whistle. The air from the bellows arrives through the conical tube at the lower end, and, escaping through a narrow slit, grazes the edge of a wedge placed opposite. A rushing noise is thus produced, which contains, among its constituents, the note to which the column of air in the pipe is capable of resounding; and as soon as this resonance occurs, the pipe speaks. Fig. 623 represents a wooden and Fig. 624 a metal organ-pipe, both of them being furnished with flute mouth-pieces. The two arrows in the sections are intended to suggest the two courses which the wind may take as it issues from the slit, one of which it actually selects to the exclusion of the other.

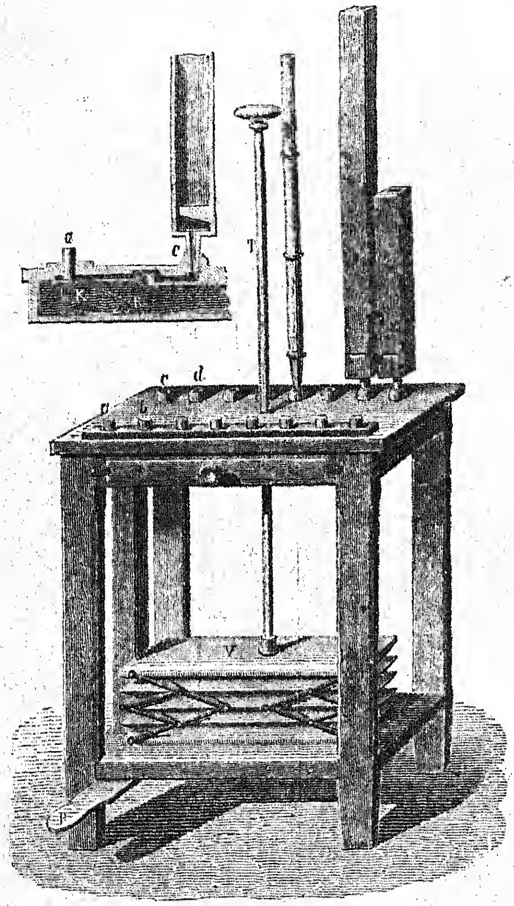


Fig. 625.—Experimental Organ.

The arrangements for admitting the wind to the pipes by putting down the keys are shown in Fig. 625. The bellows V are worked by the treadle P. The force of the blast can be increased by weighting the top of the bellows, or by pressing on the rod T. The air passes up from the bellows, through a large tube shown at one end, into a reservoir C, called the wind-chest. In the top of the wind-chest there are numerous openings *c*, *d*, &c., in which the tubes are to be fixed. The sectional drawing in the upper part of the figure shows the internal communications. A plate K, pressed up by a spring R, cuts off the tube *c* from the wind-chest, until the pin *a*

is depressed. The putting down of this pin lowers the plate, and admits the wind. This description only applies to the experimental organs which are constructed for lecture illustration. In real organs the pressure of the wind in the bellows is constant; and as this pressure would be too great for most of the pipes, the several apertures of admission are partially plugged, to diminish the force of the blast.

911. The Air is the Sonorous Body.—It is easily shown that the sound emitted by an organ-pipe depends, mainly at least, on the dimensions of the inclosed column of air, and not on the thickness or material of the pipe itself. For let three pipes, one of wood, one of copper, and the other of thick card, all of the same internal dimensions, be fixed on the wind-chest. On making them speak, it will be found that the three sounds have exactly the same pitch, and but slight difference in character. If, however, the sides of the tube are *excessively* thin, their yielding has a sensible influence, and the pitch of the sound is modified.

912. Law of Linear Dimensions.—The law of linear dimensions, stated in § 909 as applying to the vibrations of similar solid bodies, applies to gases also. Let two box-shaped pipes (Fig. 626) of precisely similar form, and having their linear dimensions in the ratio of 2 : 1, be fixed on the wind-chest; it will be found, on making them speak, that the note of the small one is an octave higher than the other;—showing double frequency of vibration.

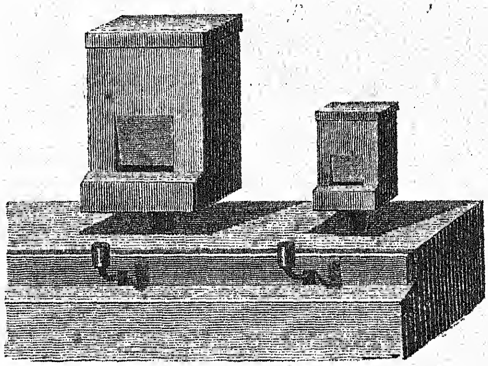


Fig. 626.—Law of Linear Dimensions.

913. Bernoulli's Laws.—The law just stated applies to the comparison of similar tubes of any shape whatever. When the length of a tube is a large multiple of its diameter, the note emitted is nearly independent of the diameter, and depends almost entirely on the length. The relations between the fundamental note of such a tube and its overtones were discovered by Daniel Bernoulli, and are as follows:—

I. *Overtones of Open Pipes.*—Let the pipe B (Fig. 627), which is

open at the upper end, be fixed on the wind-chest; let the corresponding key be put down, and the wind gradually turned on, by means of the cock below the mouth-piece. The first note heard will be feeble and deep; it is the fundamental note of the pipe. As the wind is gradually turned full on, and increasing pressure afterwards applied to the bellows, a series of notes will be heard, each higher than its predecessor. These are the overtones of the pipe. They are the harmonics of the fundamental note; that is to say, if 1 denote the frequency of vibration for the fundamental tone, the frequencies of vibration for the overtones will be approximately 2, 3, 4, 5 . . . respectively.

Fig. 627. Tubes for Overtones.

II. *Overtones of Stopped Pipes.*—If the same experiment be tried with the pipe A, which is closed at its upper end; the overtones will form the series of odd harmonics of the fundamental note, all the even harmonics being absent; in other words, the frequencies of vibration of the fundamental tone and overtones will be approximately represented by the series of odd numbers 1, 3, 5, 7 . . .

It will also be found, that if both pipes are of the same length, the fundamental note of the stopped pipe is an octave lower than that of the open pipe.

914. Mode of Production of Overtones.—In the production of the overtones, the column of air in a pipe divides itself into vibrating segments, separated by nodal cross-sections. At equal distances on opposite sides of a node, the particles of air have always equal and opposite velocities, so that the air at the node is always subjected to equal forces in opposite directions, and thus remains unmoved by their action. The portion of air constituting a vibrating segment, sways alternately in opposite directions, and as the movements in two consecutive segments are opposite, two consecutive nodes are always in opposite conditions as regards compression and extension. The middle of a vibrating segment is the place where the amplitude of vibration is greatest, and the variation of density least. It may be called an *antinode.* The distance from one node to the next is half a wave-length, and the distance from a node to an antinode is a quarter of a wave-length. Both ends of an open pipe, and the end next the mouth-piece of a stopped pipe, are antinodes, being preserved from changes of density by their free communication with

the external air. At the closed end of a stopped pipe there must always be a node.

The swaying to and fro of the internodal portions of air between fixed nodal planes, is an example of *stationary undulation;* and the vibration of a musical string is another example. A stationary undulation may always be analysed into two component undulations equal and similar to one another, and travelling in opposite directions, their common wave-length being double of the distance from node to node (§ 887). These undulations are constantly undergoing reflection from the ends of the pipe or string, and, in the case of pipes, the reflection is opposite in kind according as it takes place from a closed or an open end. In the former case a condensation propagated towards the end is reflected as a condensation, the forward-moving particles being compelled to recoil by the resistance which they there encounter; and a rarefaction is, in like manner, reflected as a rarefaction. On the other hand, when a condensation arrives at an open end, the sudden opportunity for expansion which is afforded causes an outward movement in excess of that which would suffice for equilibrium of pressure, and a rarefaction is thus produced which is propagated back through the tube. A condensation is thus reflected as a rarefaction; and a rarefaction is, in like manner, reflected as a condensation.

The period of vibration of the fundamental note of a stopped pipe is the time required for propagating a pulse through four times the length of the pipe. For let a condensation be suddenly produced at the lower end by the action of the vibrating lip. It will be propagated to the closed end and reflected back, thus travelling over twice the length of the pipe. On arriving at the aperture where the lip is situated, it is reflected as a rarefaction. This rarefaction travels to the closed end and back, as the condensation did before it, and is then reflected from the aperture as a condensation. Things are now in their initial condition, and one complete vibration has been performed. The period of the movements of the lip is determined by the arrival of these alternate condensations and rarefactions; and the lip, in its turn, serves to divert a portion of the energy of the blast, and employ it in maintaining the energy of the vibrating column.

The wave-length of the fundamental note of a stopped pipe is thus four times the length of the pipe.

In an open pipe, a condensation, starting from the mouth-piece, is reflected from the other end as a rarefaction. This rarefaction, on

open at the upper end, be fixed on the wind-chest; let the corresponding key be put down, and the wind gradually turned on, by means of the cock below the mouth-piece. The first note heard will be feeble and deep; it is the fundamental note of the pipe. As the wind is gradually turned full on, and increasing pressure afterwards applied to the bellows, a series of notes will be heard, each higher than its predecessor. These are the overtones of the pipe. They are the harmonics of the fundamental note; that is to say, if 1 denote the frequency of vibration for the fundamental tone, the frequencies of vibration for the overtones will be approximately 2, 3, 4, 5 . . . respectively.

Fig. 627. Tubes for Overtones.

II. *Overtones of Stopped Pipes.*—If the same experiment be tried with the pipe A, which is closed at its upper end; the overtones will form the series of odd harmonics of the fundamental note, all the even harmonics being absent; in other words, the frequencies of vibration of the fundamental tone and overtones will be approximately represented by the series of odd numbers 1, 3, 5, 7 . . .

It will also be found, that if both pipes are of the same length, the fundamental note of the stopped pipe is an octave lower than that of the open pipe.

914. Mode of Production of Overtones.—In the production of the overtones, the column of air in a pipe divides itself into vibrating segments, separated by nodal cross-sections. At equal distances on opposite sides of a node, the particles of air have always equal and opposite velocities, so that the air at the node is always subjected to equal forces in opposite directions, and thus remains unmoved by their action. The portion of air constituting a vibrating segment, sways alternately in opposite directions, and as the movements in two consecutive segments are opposite, two consecutive nodes are always in opposite conditions as regards compression and extension. The middle of a vibrating segment is the place where the amplitude of vibration is greatest, and the variation of density least. It may be called an *antinode*. The distance from one node to the next is half a wave-length, and the distance from a node to an antinode is a quarter of a wave-length. Both ends of an open pipe, and the end next the mouth-piece of a stopped pipe, are antinodes, being preserved from changes of density by their free communication with

the external air. At the closed end of a stopped pipe there must always be a node.

The swaying to and fro of the internodal portions of air between fixed nodal planes, is an example of *stationary undulation;* and the vibration of a musical string is another example. A stationary undulation may always be analysed into two component undulations equal and similar to one another, and travelling in opposite directions, their common wave-length being double of the distance from node to node (§ 887). These undulations are constantly undergoing reflection from the ends of the pipe or string, and, in the case of pipes, the reflection is opposite in kind according as it takes place from a closed or an open end. In the former case a condensation propagated towards the end is reflected as a condensation, the forward-moving particles being compelled to recoil by the resistance which they there encounter; and a rarefaction is, in like manner, reflected as a rarefaction. On the other hand, when a condensation arrives at an open end, the sudden opportunity for expansion which is afforded causes an outward movement in excess of that which would suffice for equilibrium of pressure, and a rarefaction is thus produced which is propagated back through the tube. A condensation is thus reflected as a rarefaction; and a rarefaction is, in like manner, reflected as a condensation.

The period of vibration of the fundamental note of a stopped pipe is the time required for propagating a pulse through four times the length of the pipe. For let a condensation be suddenly produced at the lower end by the action of the vibrating lip. It will be propagated to the closed end and reflected back, thus travelling over twice the length of the pipe. On arriving at the aperture where the lip is situated, it is reflected as a rarefaction. This rarefaction travels to the closed end and back, as the condensation did before it, and is then reflected from the aperture as a condensation. Things are now in their initial condition, and one complete vibration has been performed. The period of the movements of the lip is determined by the arrival of these alternate condensations and rarefactions; and the lip, in its turn, serves to divert a portion of the energy of the blast, and employ it in maintaining the energy of the vibrating column.

The wave-length of the fundamental note of a stopped pipe is thus four times the length of the pipe.

In an open pipe, a condensation, starting from the mouth-piece, is reflected from the other end as a rarefaction. This rarefaction, on

reaching the mouth-piece, is reflected as a condensation; and things are thus in their initial state after the length of the pipe has been traversed twice. The period of vibration of the fundamental note is accordingly the time of travelling over twice the length of the pipe; and its wave-length is twice the length of the pipe. In every case of longitudinal vibration, if the reflection is alike at both ends, the wave-length of the fundamental tone is twice the distance between the ends.

915. **Explanation of Bernoulli's Laws.**—In investigating the theoretical relations between the fundamental tone and overtones for a pipe of either kind, it is convenient to bear in mind that the distance from an open end to the nearest node is a quarter of a wave-length of the note emitted.

In the case of the open pipe the first or fundamental tone requires one node, which is at the middle of the length. The second tone requires two nodes, with half a wave-length between them, while each of them is a quarter of a wave-length from the nearest end. A quarter wave-length has thus only half the length which it had for the fundamental tone, and the frequency of vibration is therefore doubled.

The third tone requires three nodes, and the distance from either end to the nearest node is $\frac{1}{6}$ of the length of the pipe, instead of $\frac{1}{2}$ the length as in the case of the first tone. The wave-length is thus divided by 3, and the frequency of vibration is increased threefold. We can evidently account in this way for the production of the complete series of harmonics of the fundamental note.

In the case of the stopped pipe, the mouth-piece is always distant a quarter wave-length from the nearest node, and this must be distant an even number of quarter wave-lengths from the stopped end, which is itself a node.

For the fundamental tone, a quarter wave-length is the whole length of the pipe.

For the second tone, there is one node besides that at the closed end, and its distance from the open end is $\frac{1}{3}$ of the length of the pipe.

For the third tone, there are two nodes besides that at the closed end. The distance from the open end to the nearest node is therefore $\frac{1}{5}$ of the length of the pipe.

The wave-lengths of the successive tones, beginning with the fundamental, are therefore as $1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7} \ldots$, and their vibration-frequencies are as $1, 3, 5, 7 \ldots$

Also, since the wave-length of the fundamental tone is four times the length of the pipe if stopped, and only twice its length if open, it is obvious that the wave-length is halved, and the frequency of vibration doubled, by unstopping the pipe.

No change of pitch, or only very slight change, will be produced by inserting a solid partition at a node, or by putting an antinode in free communication with the external air. These principles can be illustrated by means of the jointed pipe represented in Fig. 628.

Fig. 628. Jointed Pipe.

916. Application to Rods and Strings.—The same laws which apply to open organ-pipes, also apply to the longitudinal vibrations of rods free at both ends, and to both the longitudinal and transverse vibrations of strings. In all these cases the overtones form the complete series of harmonics of the first or fundamental tone, and the period of vibration for this first tone is the time occupied by a pulse in travelling over twice the length of the given rod or string. In the case of longitudinal vibrations the velocity of a pulse is $\sqrt{\frac{M}{D}}$, M denoting the value of Young's modulus for the rod or string, and D its density. This is identical with the velocity of sound through the rod or string, and is independent of its tension. In the case of transverse pulses in a string (regarded as perfectly flexible), the formula for the velocity of transmission (1) § 901, may be written $\sqrt{\frac{F}{D}}$, F denoting the stretching force per unit of sectional area. The ratio of the latter velocity to the former is $\sqrt{\frac{F}{M}}$, which is always a small fraction, since $\frac{F}{M}$ expresses the fraction of itself by which the string is lengthened by the force F.

If a rod, free at both ends, is made to vibrate longitudinally, its nodes and antinodes will be distributed exactly in the same way as those of an open organ-pipe. The experiment can be performed by holding the rod at a node, and rubbing it with rosined chamois leather.

917. Application to Measurement of Velocity in Gases.—Let v denote the velocity of sound in a particular gas, in feet per second, λ the wave-length of a particular note in this gas in feet, and n the frequency of vibration for this note, that is the number of vibrations

per second which produce it. Then λ is the distance travelled in $\frac{1}{n}$ of a second, and the distance travelled in a second is

$$v = n\lambda.$$

For the same note, n is constant for all media whatever, and v varies directly as λ. The velocities of sound in two gases may thus be compared by observing the lengths of vibrating columns of the two gases which give the same note; or if columns of equal length be employed, the velocities will be directly as the frequencies of vibration, which are determined by observing the pitch of the notes emitted.

By these methods, Dulong, and more recently Wertheim, have determined the velocity of sound in several different gases. The following are Wertheim's results, in metres per second, the gases being supposed to be at 0° C.

Air,	331	Carbonic acid,	262
Oxygen,	317	Nitrous oxide,	262
Hydrogen,	1269	Olefiant gas,	314
Carbonic oxide,	337		

The same principle is applicable to liquids and solids; and it was by means of the longitudinal vibrations of rods that the velocities given in § 880 were ascertained.

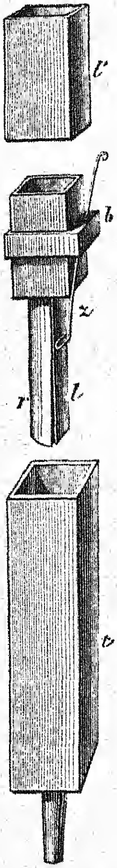


Fig. 629.—Reed Pipe.

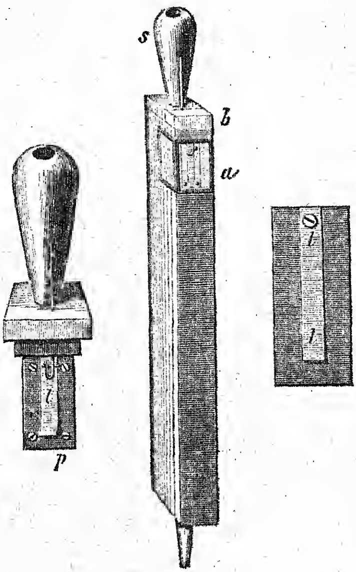


Fig. 630.—Free Reed.

918. Reed-pipes.—Instead of the flute mouth-piece above described, organ-pipes are often furnished with what is called a *reed*. A reed contains an elastic plate l (Figs. 629, 630) called the *tongue*, which, by its vibrations, alternately opens and closes or nearly closes an aperture through which the wind passes. In Fig. 629, the air from the bellows enters first

the lower part t of the pipe, and thence (when permitted by the tongue) passes through the channel[1] r into the upper part t'. The stiff wire z, movable with considerable friction through the hole b, limits the vibrating portion of the tongue, and is employed for tuning. Reed-pipes are often terminated above by a trumpet-shaped expansion.

A *striking reed* (Fig. 629) is one whose tongue closes the aperture by covering it. The tongue should be so shaped as not to strike along its whole length at once, but to roll itself down over the aperture. In the *free reed* (Fig. 630) the tongue can pass completely through.

The striking reed is generally preferred in organs, its peculiar character rendering it very effective by way of contrast. It is always used for the *trumpet* stop. Reed-pipes can be very strongly blown without breaking into overtones. Their pitch, however, if they are of the striking kind, is not independent of the pressure of the wind, but gradually rises as the pressure increases. Free reeds, which are used for harmoniums, accordions, and concertinas, do not change in pitch with change of pressure.

Elevation of temperature sharpens pipes with flute mouth-pieces, and flattens reed-pipes. The sharpening is due to the increased velocity of sound in hot air. The flattening is due to the diminished elasticity of the metal tongue. It is thus proved that the pitch of a reed-pipe is not always that due to the free vibration of the inclosed air, but may be modified by the action of the tongue.

919. Wind-instruments.—In all wind-instruments, the sound is originated by one of the two methods just described. With the flute-pipe must be classed the flute, the flageolet, and the Pandean-pipes. The clarionet, hautboy, and bassoon have reed mouth-pieces, the vibrating tongue being a piece of reed or cane. In the bugle, trumpet, and French-horn, which are mere tubes without keys, the lips of the performer act as the reed-tongue, and the notes produced are approximately the natural overtones. These, when of high order, are so near together, that a gamut can be formed by properly selecting from among them.

The fingering of the flute and clarionet, has the effect sometimes of altering the effective length of the vibrating column of air, and sometimes of determining the production of overtones. In the

[1] The piece r, which is approximately a half cylinder, is called the *reed* by organ-builders.

trombone and cornet-à-piston, the length of the vibrating column of air is altered. The harmonium, accordion, and concertina are reed instruments, the reeds employed being always of the free kind.

920. Manometric Flames.—Koenig, of Paris, constructs several forms of apparatus, in which the variations of pressure produced by vibrations of air in a pipe are rendered evident to the eye by their effect upon flames. One of these is represented in Fig. 631. Three small gas-burners are fixed at definite points in the side of a pipe, as represented in the figure. When the pipe gives its second tone, the central flame is at an antinode and remains unaffected, while the other two, being at nodes, are agitated or blown out. When it gives its first tone, the central flame, which is now at a node, is more powerfully affected than the others. The gas which supplies these burners is separated from the air in the pipe only by a thin membrane. When the pipe is made to speak, the flame at the node is violently agitated, in consequence of the changes of pressure on the back of the membrane, while those at the ventral points are scarcely affected. The agitation of the flame is a true vibration; and, when examined by the aid of a revolving mirror, presents the appearance of tongues of flame alternating with nearly dark spaces. If two pipes, one an octave higher than the other, are connected with the same gas flame, or with two gas flames which can be viewed in the same mirror, the tongues of flame corresponding to the upper octave are seen to be twice as numerous as the others.

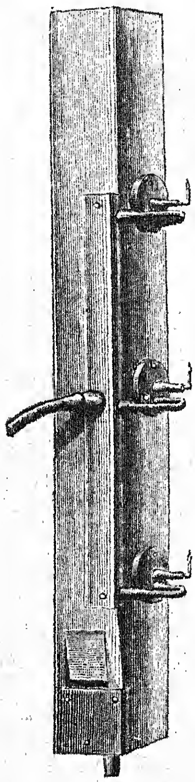


Fig. 631.—Manometric Flames.

CHAPTER LXV.

ANALYSIS OF VIBRATIONS. CONSTITUTION OF SOUNDS.

921. Optical Examination of Sonorous Vibrations.—Sound is a special sensation belonging to the sense of hearing; but the vibrations which are its physical cause often manifest themselves to other senses. For instance, we can often feel the tremors of a sonorous body by touching it; we see the movements of the sand on a vibrating plate, the curve traced by the style of a vibroscope, &c. The aid which one sense can thus furnish in what seems the peculiar province of another is extremely interesting. M. Lissajous has devised a very beautiful optical method of examining sonorous vibrations, which we will briefly describe. Omit

922. Lissajous' Experiment.—Suppose we introduce into a dark room (Fig. 632) a beam of solar rays, which, after passing through a lens L, is reflected, first, from a small mirror fixed on one of the branches of a tuning-fork D, and then from a second mirror M, which throws it on a screen E; we can thus, by proper adjustments, form upon the screen a sharp and bright image of the sun, which will appear as a small spot of light. As long as the apparatus remains at rest, we shall not observe any movement of the image; but if the tuning-fork vibrates, the image will move rapidly up and down along the line I, I′, producing, in consequence of the persistence of impressions, the appearance of a vertical line of light. If the tuning-fork remains at rest, but the mirror M is rotated through a small angle about a vertical axis, the image will move horizontally. Consequently, if both these motions take place simultaneousely, the spot of light will trace out on the screen a sinuous line, as represented in the figure, each S-shaped portion corresponding to one vibration of the tuning-fork.

Now, let the mirror M be replaced by a small mirror attached to

a second tuning-fork, which vibrates in a horizontal plane, as in

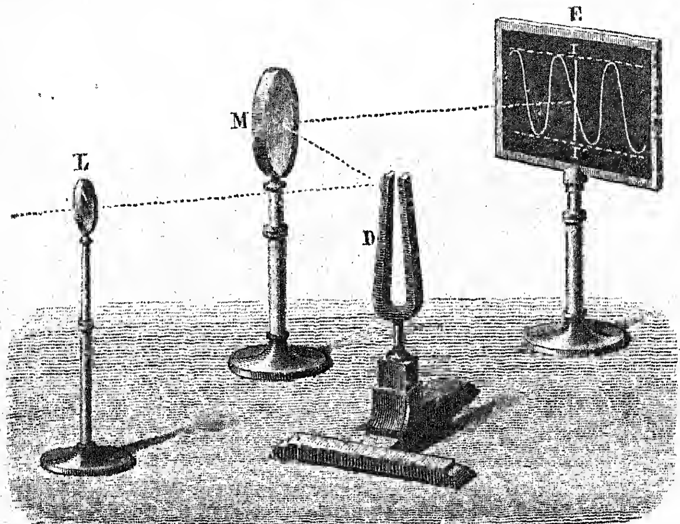


Fig. 632.—Principle of Lissajous' Experiment.

Fig. 633. If this fork vibrates alone, the image will move to and

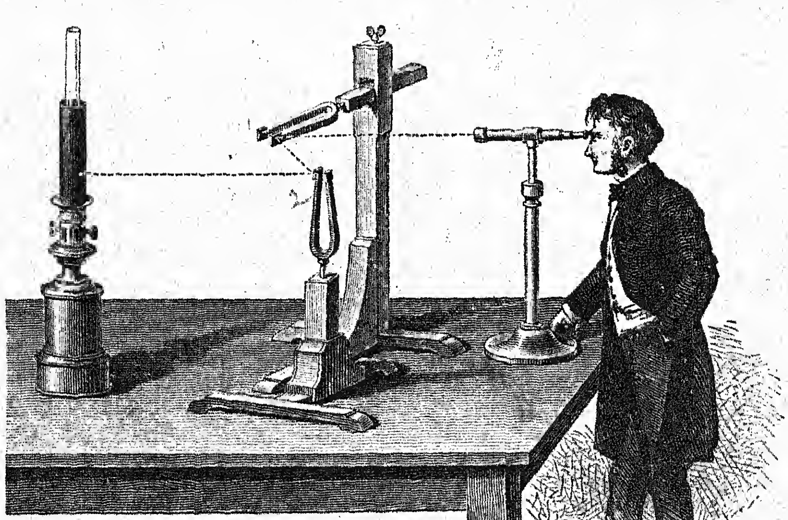


Fig. 633.—Lissajous' Experiment.

fro horizontally, presenting the appearance of a horizontal line of

light, which gradually shortens as the vibrations die away. If both forks vibrate simultaneously, the spot of light will rise and fall according to the movements of the first fork, and will travel left and right according to the movements of the second fork. The curve actually described, as the resultant of these two component motions, is often extremely beautiful. Some varieties of it are represented in Fig. 634.

Instead of throwing the curves on a screen, we may see them by looking into the second mirror, either with a telescope, as in Fig. 633,

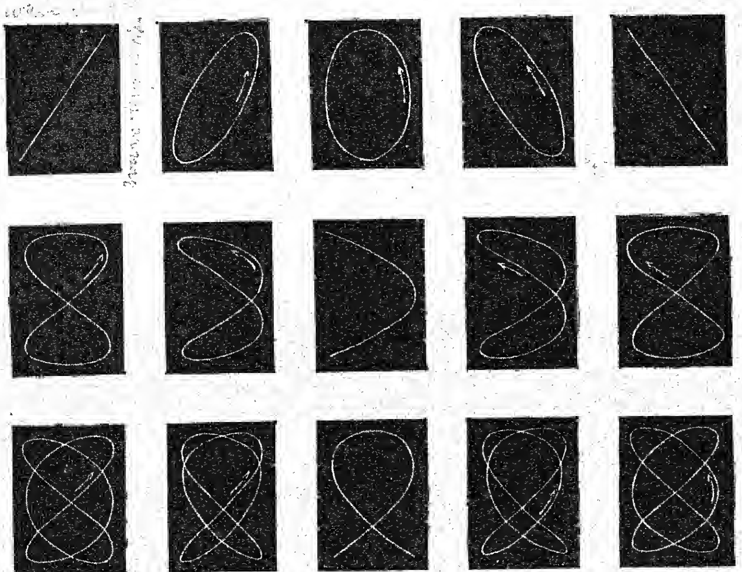


Fig. 634.—Lissajous' Figures, Unison, Octave, and Fifth.

or with the naked eye. In this form of the experiment, a lamp surrounded by an opaque cylinder, pierced with a small hole just opposite the flame, as represented in the figure, is a very convenient source of light.

The movement of the image depends almost entirely on the angular movements of the mirrors, not on their movements of translation; but the distinction is of no importance, for, in the case of such small movements, the linear and angular changes may be regarded as strictly proportional.

Either fork vibrating alone would cause the image to execute *simple harmonic motion* (§§ 109–111), or, as it may conveniently

be called, *simple vibration;* so that the movement actually executed will be the resultant of two simple harmonic motions in directions perpendicular to each other.

Suppose the two forks to be in unison. Then the two simple harmonic motions will have the same period, and the path described will always be some kind of ellipse,[1] the circle and straight line being included as particular cases. It will be a straight line if both forks pass through their positions of equilibrium at the same instant. In order that it may be a circle, the amplitudes of the two simple harmonic motions must be equal, and one fork must be in a position of maximum displacement when the other is in the position of equilibrium.

If the unison were rigorous, the curve once obtained would remain unchanged, except in so far as its breadth and height became reduced by the dying away of the vibrations. But this perfect unison is never attained in practice, and the eye detects changes depending on differences of pitch too minute to be perceived by the ear. These changes are illustrated by the upper row of forms in Fig. 634, commencing, say, with the sloping straight line at the left hand, which gradually opens out into an ellipse, and afterwards contracts into a straight line, sloping the opposite way. It then retraces its steps, the motion being now in opposition to the arrows in the figure, and then repeats the same changes.

If the interval between the two forks is an octave, we shall obtain the curves represented in the second row;[2] if the interval is a fifth, we shall obtain the curves in the lowest row. In each case the order of the changes will be understood by proceeding from left to right,

[1] Employing horizontal and vertical co-ordinates, and denoting the amplitudes by a and b, we have, in the case of unison, $\frac{x}{a} = \sin\theta$, $\frac{y}{b} = \sin(\theta + \beta)$, where β denotes the difference of phase, and θ is an angle varying directly as the time. Eliminating θ, we obtain the equation to an ellipse, whose form and dimensions depend upon the given quantities, a, b, β.

[2] The middle curve in this row is a parabola, and corresponds to the elimination of θ between the equations $\frac{x}{a} = \cos 2\theta$, $\frac{y}{a} = \cos\theta$. The coefficient 2 indicates the double frequency of horizontal as compared with vertical vibrations.

The general equations to Lissajous' figures are $\frac{x}{a} = \sin m\theta$, $\frac{y}{b} = \sin(n\theta + \beta)$, where m and n are proportional to the frequencies of horizontal and vertical vibrations. The gradual changes from one figure to another depend on the gradual change of β, and all the figures can be inscribed in a rectangle, whose length and breadth are $2a$ and $2b$.

and then back again; but the curves obtained in returning will be inverted.

923. Optical Tuning.—By the aid of these principles, tuning-forks can be compared with a standard fork with much greater precision than would be attainable by ear. Fig. 635 represents a convenient

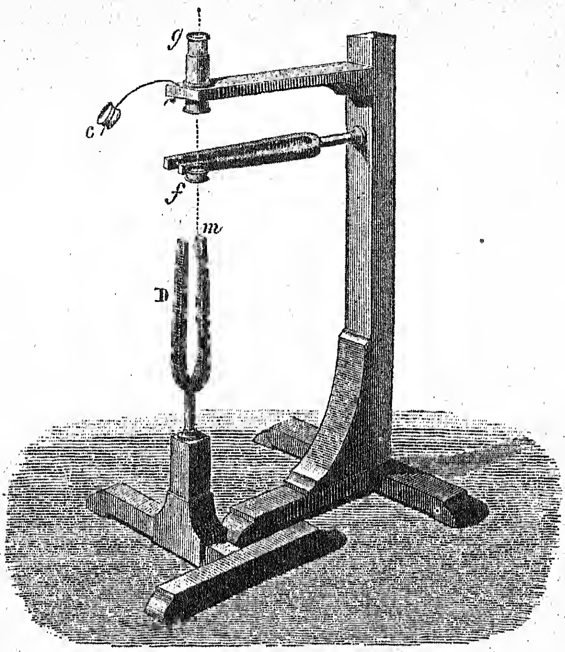


Fig. 635.—Optical Comparison of Tuning-forks.

arrangement for this purpose. A lens *f* is attached to one of the prongs of a standard fork, which vibrates in a horizontal plane; and above it is fixed an eye-piece *g*, the combination of the two being equivalent to a microscope. The fork to be compared is placed upright beneath, and vibrates in a vertical plane, the end of one prong being in the focus of the microscope. A bright point *m*, produced by making a little scratch on the end of the prong with a diamond, is observed through the microscope, and is illuminated, if necessary, by converging a beam of light upon it through the lens *c*. When the forks are set vibrating, the bright point is seen as a luminous ellipse, whose permanence of form is a test of the closeness of the unison. The ellipse will go through a complete cycle of changes in the time required for one fork to gain a complete vibration on the other.

924. Other Modes of producing Lissajous' Figures.—An arrangement devised in 1844 by Professor Blackburn, of Glasgow, then a student at Cambridge, affords a very easy mode of obtaining, by a slow motion, the same series of curves which, in the above arrangements, are obtained by a motion too quick for the eye to follow. A cord A B C (Fig. 636) is fastened at A and C, leaving more or less

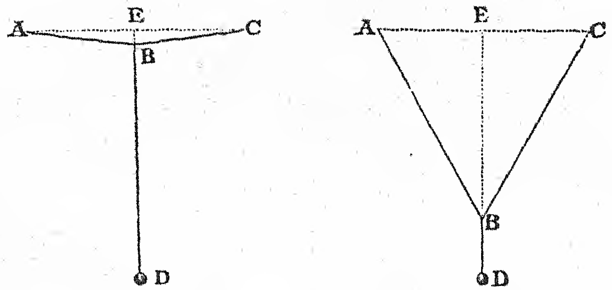


Fig. 636.—Blackburn's Pendulum.

slack, according to the curves which it is desired to obtain; and to any intermediate point B of the cord another string is tied, carrying at its lower end a heavy body D to serve as pendulum-bob.

If, when the system is in equilibrium, the bob is drawn aside in the plane of A B C and let go, it will execute vibrations in that plane, the point B remaining stationary, so that the length of the pendulum is B D. If, on the other hand, it be drawn aside in a plane perpendicular to the plane A B C, it will vibrate in this perpendicular plane, carrying the whole of the string with it in its motion, so that the length of the pendulum is the distance of the bob from the point E, in which the straight line A C is cut by D B produced. The frequencies of vibration in the two cases will be inversely as the square roots of the pendulum-lengths B D, E D.

If the bob is drawn aside in any other direction, it will not vibrate in one plane, but will perform movements compounded of the two independent modes of vibration just described, and will thus describe curves identical with Lissajous'. If the ratio of E D to B D is nearly equal to unity, as in the left-hand figure, we shall have curves corresponding to approximate unison. If it be approximately 4, as in the right-hand figure, we shall obtain the curves of the octave. Traces of the curves can be obtained by employing for the bob a

vessel containing sand, which runs out through a funnel-shaped opening at the bottom.[1]

The curves can also be exhibited by fixing a straight elastic rod at one end, and causing the other end to vibrate transversely. This was the earliest known method of obtaining them. If the flexural rigidity of the rod is precisely the same for all transverse directions, the vibrations will be executed in one plane; but if there be any inequality in this respect, there will be two mutually perpendicular directions possessing the same properties as the two principal directions of vibration in Blackburn's pendulum. A small bright metal knob is usually fixed on the vibrating extremity to render its path visible. The instrument constructed for this mode of exhibiting the figures is called a *kaleidophone.* In its best form (devised by Professor Barrett) the upper and lower halves of the rod (which is vertical) are flat pieces of steel, with their planes at right angles, and a stand is provided for clamping the lower piece at any point of its length that may be desired, so as to obtain any required combination.

925. Character.—*Character* or *timbre,* which we have already defined in § 889, must of necessity depend on the *form* of the vibration of the aerial particles by which sound is transmitted, the word *form* being used in the metaphorical sense there explained, for in the literal sense the form is always a straight line. When the changes of density are represented by ordinates of a curve, as in Fig. 603, the form of this curve is what is meant by the form of vibration.

The subject of *timbre* has been very thoroughly investigated in recent years by Helmholtz; and the results at which he has arrived are now generally accepted as correct.

The first essential of a musical note is, that the aerial movements which constitute it shall be strictly *periodic;* that is to say, that each vibration shall be exactly like its successor, or at all events, that, if there be any deviation from strict periodicity, it shall be so gradual as not to produce sensible dissimilarity between several consecutive vibrations of the same particle.

There is scarcely any proposition more important in its application

[1] Mr. Hubert Airy has obtained very beautiful traces by attaching a glass pen to the bob (see *Nature,* Aug. 17 and Sept. 7, 1871), and in Tisley's *harmonograph* the same result is obtained by means of two pendulums, one of which moves the paper and the other the pen.

to modern physical investigations than the following mathematical theorem, which was discovered by Fourier:—*Any periodic vibration executed in one line can be definitely resolved into simple vibrations, of which one has the same frequency as the given vibration, and the others have frequencies* 2, 3, 4, 5 . . . *times as great*, no fractional multiples being admissible. The theorem may be briefly expressed by saying that *every periodic vibration consists of a fundamental simple vibration and its harmonics.*

We cannot but associate this mathematical law with the experimental fact, that a trained ear can detect the presence of harmonics in all but the very simplest musical notes. The analysis which Fourier's theorem indicates, appears to be actually performed by the auditory apparatus.

The *constitution* of a periodic vibration may be said to be known if we know the ratios of the amplitudes of the simple vibrations which compose it; and in like manner the constitution of a sound may be said to be known if we know the relative intensities of the different elementary tones which compose it.

Helmholtz infers from his experiments that the *character* of a musical note depends upon its *constitution* as thus defined; and that, while change of intensity in any of the components produces a modification of character, change of phase has no influence upon it whatever. Sir W. Thomson, in a paper "On Beats of Imperfect Harmonics,"[1] adduces strong evidence to show that change of phase has, in some cases at least, an influence on character.

The harmonics which are present in a note, usually find their origin in the vibrations of the musical instrument itself. In the case of stringed instruments, for example, along with the vibration of the string as a whole, a number of segmental vibrations are simultaneously going on. Fig. 637 represents curves obtained by the composition of the fundamental mode of vibration with another an octave higher. The broken lines indicate the forms which the string would assume if yielding only its fundamental note.[2] The continuous lines in the first and third figures are forms which a string may assume in its two positions of greatest displacement, when yielding the octave along with the fundamental, the time required for the

[1] *Proc. R. S. E.* 1878.

[2] The form of a string vibrating so as to give only one tone (whether fundamental or harmonic) is a curve of sines, all its ordinates increasing or diminishing in the same proportion, as the string moves.

string to pass from one of these positions to the other being the same as the time in which each of its two segments moves across and back

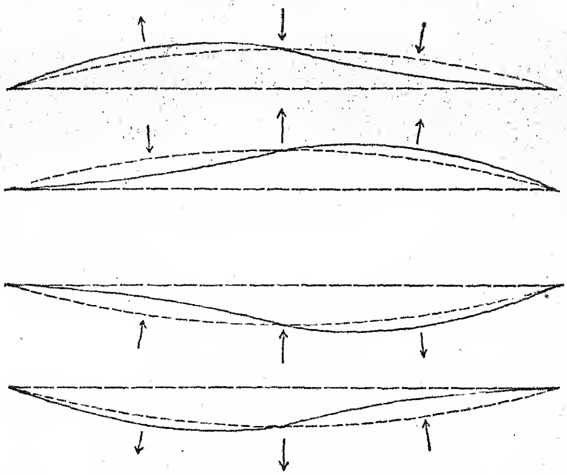


Fig. 637.—String giving first Two Tones.

again. The second and fourth figures must in like manner be taken together, as representing a pair of extreme positions. The number of harmonics thus yielded by a pianoforte wire is usually some four or five; and a still larger number are yielded by the strings of a violin.

The notes emitted from wide organ-pipes with flute mouth-pieces are very deficient in harmonics. This defect is remedied by combining with each of the larger pipes a series of smaller pipes,[1] each yielding one of its harmonics. An ordinary listener hears only one note, of the same pitch as the fundamental, but much richer in character than that which the fundamental pipe yields alone. A trained ear can recognize the individual harmonics in this case as in any other.

[1] The stops called *open diapason* and *stop diapason* (consisting respectively of open and stopped pipes), give the fundamental tone, almost free from harmonics. The stop absurdly called *principal* gives the second tone, that is the octave above the fundamental. The stops called *twelfth* and *fifteenth* give the third and fourth tones, which are a twelfth (octave + fifth), and a fifteenth (double octave) above the fundamental. The fifth, sixth, and eighth tones are combined to form the stop called *mixture*.

As many of our readers will be unacquainted with the structure of organs, it may be desirable to state that an organ contains a number of complete instruments, each consisting of several octaves of pipes. Each of these complete instruments is called a *stop*, and is brought into use at the pleasure of the organist by pulling out a slide, by means of a knob-handle, on which the name of the stop is marked. To throw it out of use, he pushes in the slide. A large number of stops are often in use at once.

It is important to remark, that though the presence of harmonic subdivisions in a vibrating body necessarily produces harmonics in the sound emitted, the converse cannot be asserted. Simple vibrations, executed by a vibrating body, produce vibrations *of the same frequency* as their own, in any medium to which they are transmitted, but not necessarily *simple* vibrations. If they produce compound vibrations, these, as we have seen (§ 925), must consist of a fundamental simple vibration and its harmonics.

926. Helmholtz's Resonators.—Helmholtz derived material aid in his researches from an instrument devised by himself, and called a *resonator* or *resonance globe* (Fig. 638). It is a hollow globe of thin

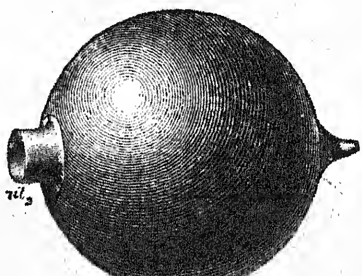


Fig. 638.—Resonator.

brass, with an opening at each end, the larger one serving for the admission of sound, while the smaller one is introduced into the ear. The inclosed mass of air has, like the column of air in an organ-pipe, a particular fundamental note of its own, depending upon its size; and whenever a note of this particular pitch is sounded in its neighbourhood, the inclosed air takes it up and intensifies it by resonance. In order to test the presence or absence of a particular harmonic in a given musical tone, a resonator, in unison with this harmonic, is applied to the ear, and if the resonator speaks it is known that the harmonic is present. These instruments are commonly constructed so as to form a series, whose notes correspond to the bass C of a man's voice, and its successive harmonics as far as the 10th or 12th.

Koenig has applied the principle of manometric flames to enable a large number of persons to witness the analysis of sounds by resonators. A series of 6 resonators, whose notes have frequencies proportional to 1, 2, 3, 4, 5, 6, are fixed on a stand (Fig. 639), and their smaller ends, instead of being applied to the ear, are connected each

with a separate manometric capsule, which acts on a gas jet. When the mirrors are turned, it is easy to see which of the flames vibrate while a sonorous body is passed in front of the resonators.

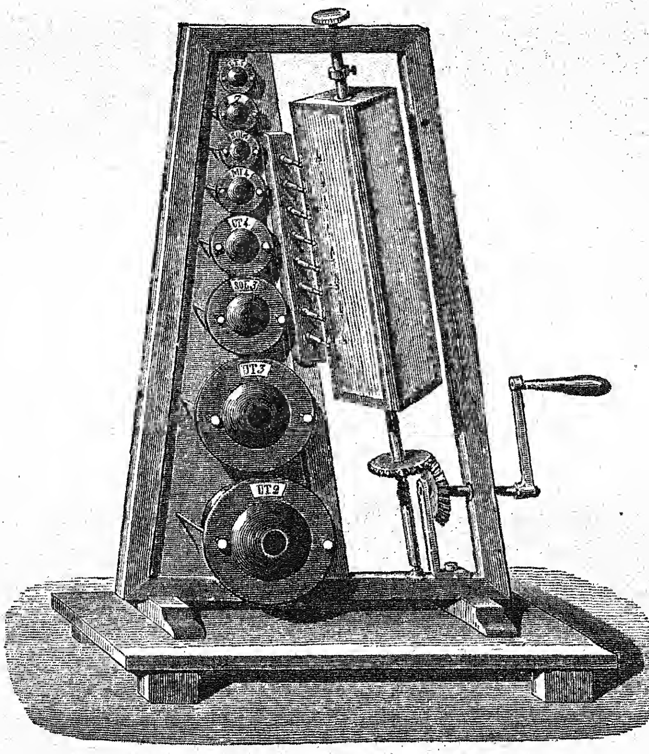


Fig. 639.—Analysis by Manometric Flames.

A simple tone, unaccompanied by harmonics, is dull and uninteresting, and, if of low pitch, is very destitute of penetrating quality.

Sounds composed of the first six elementary tones in fair proportion, are rich and sweet.

The higher harmonics, if sufficiently subdued, may also be present without sensible detriment to sweetness, and are useful as contributing to expression. When too loud, they render a sound harsh and grating; an effect which is easily explained by the discordant combinations which they form one with another; the 8th and 9th tones, for example, are at the same interval as the notes *Do* and *Re*.

927. Vowel Sounds.—The human voice is extremely rich in harmonics, as may be proved by applying the series of resonators to the

ear while the fundamental note is sung. The origin of the tones of the voice is in the vocal chords, which, when in use, form a diaphragm with a slit along its middle. The edges of this slit vibrate when air is forced through, and, by alternately opening and closing the passage, perform the part of a reed. The cavity of the mouth serves as a resonance chamber, and reinforces particular notes depending on the position of the organs of speech. It is by this resonance that the various vowel sounds are produced. The deepest pitch belongs to the vowel sound which is expressed in English by *oo* (as in *moon*), and the highest to *ee* (as in *screech*).

Willis in 1828[1] succeeded in producing the principal vowel sounds by a single reed fitted to various lengths of tube. Wheatstone, a few years later, made some advances in theory,[2] and constructed a machine by which nearly all articulate sounds could be imitated.

Excellent imitations of some of the vowel sounds can be obtained by placing Helmholtz's resonators, one at a time, on a free-reed pipe, the small end of the resonator being inserted in the hole at the top of the pipe.

The best determinations of the particular notes which are reinforced in the case of the several vowel sounds, have been made by Helmholtz, who employed several methods, but chiefly the two following:—

1. Holding resonators to the ear, while a particular vowel sound was loudly sung.

2. Holding vibrating tuning-forks in front of the mouth when in the proper position for pronouncing a given vowel; and observing which of them had their sounds reinforced by resonance.[3]

Helmholtz has verified his determinations synthetically. He employs a set of tuning-forks which are kept in vibration by the alternate making and unmaking of electro-magnets, the circuit being made and broken by the vibrations of one large fork of 64 vibrations per second. The notes of the other forks are the successive harmonics of this fundamental note. Each fork is accompanied by a

[1] *Cambridge Transactions*, vol. iii.

[2] *London and Westminster Review*, October, 1837.

[3] According to Koenig (*Comptes Rendus*, 1870) the notes of strongest resonance for the vowels *u, o, a, e, i*, as pronounced in North Germany, are the five successive octaves of B flat, commencing with that which corresponds to the space above the top line of the base clef. Willis, Helmholtz, and Koenig all agree as regards the note of the vowel *o*, which is very nearly that of a common A tuning-fork. They are also agreed respecting the note of *a* (as in *father*), which is an octave higher.

resonance-tube, which, when open, renders the note of the fork audible at a distance; and by means of a set of keys, like those of a piano, any of these tubes can be opened at pleasure. The different vowel-sounds can thus be produced by employing the proper combinations.

The same apparatus served for establishing the principle (§ 925), that the character of a musical sound depends only on *constitution*, irrespective of change of phase.

928. Phonograph.—Mr. Edison of New York has been successful in constructing an instrument which can reproduce articulate sounds spoken into it. The voice of the speaker is directed into a funnel, which converges the sonorous waves upon a diaphragm carrying a style. The vibrations of the diaphragm are impressed by means of this style upon a sheet of tin-foil, which is fixed on the outside of a cylinder to which a spiral motion is given as in the vibroscope (Fig. 616). After this has been done, the cylinder with the tin-foil on it is shifted back to its original position, the style is brought into contact

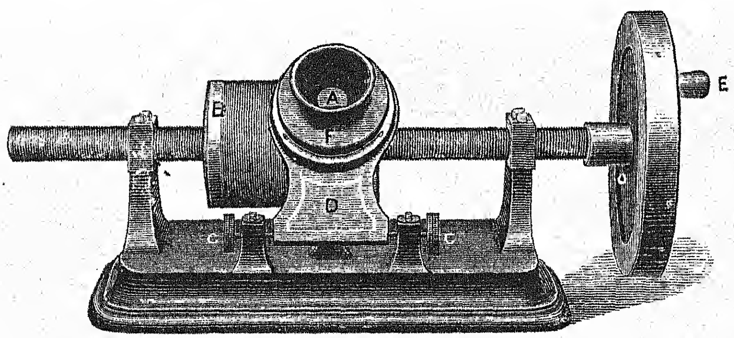


Fig. 640.—Phonograph.

with the tin-foil as at first, and the cylinder is then turned as before. The indented record is thus passed beneath the style, and forces it and the attached diaphragm to execute movements resembling their original movements. The diaphragm accordingly emits sounds which are imitations of those previously spoken to it. Tunes sung into the funnel are thus reproduced with great fidelity, and sentences clearly spoken into it are reproduced with sufficient distinctness to be understood.

The instrument is represented in Fig. 640. By turning the handle E, which is attached to a massive fly-wheel, the cylinder B is made

to revolve and at the same time to travel longitudinally, as the axle on which it is mounted is a screw working in a fixed nut. The surface of the cylinder is also fluted screw-fashion, the distance between its flutings being the same as the distance between the threads on the axle. A is the diaphragm, of thin sheet-iron, having the style fixed to its centre but not visible in the figure. The diaphragm and funnel are carried by the frame D, which turns on a hinge at the bottom. C C are adjusting screws for bringing the style exactly opposite the centre of the groove on the cylinder, and another screw is provided beneath the frame D, for making the style project so far as to indent the tin-foil without piercing it. The tin-foil is put round the cylinder, and lightly fastened with cement, so that it can be quickly taken off and changed.

In another form of the instrument, the rotation of the cylinder is effected by means of a driving weight and governor, which give it a constant velocity. This is a great advantage in reproducing music, but is of little or no benefit for speech.

CHAPTER LXVI.

CONSONANCE, DISSONANCE, AND RESULTANT TONES.

929. **Concord and Discord.**—Every one not utterly destitute of musical ear is familiar with the fact that certain notes, when sounded together, produce a pleasing effect by their combination, while certain others produce an unpleasing effect. The combination of two or more notes, when agreeable, is called *concord* or *consonance;* when disagreeable, *discord* or *dissonance*. The distinction is found to depend almost entirely on difference of pitch, that is, on relative frequency of vibration; so that the epithets consonant and dissonant can with propriety be applied to intervals.

The following intervals are consonant: unison (1 : 1), octave (1 : 2), octave + fifth (1 : 3), double octave (1 : 4), fifth (2 : 3), fourth (3 : 4).

The major third (4 : 5) and major sixth (3 : 5), together with the minor third (5 : 6) and minor sixth (5 : 8), are less perfect in their consonance.

The second and the seventh, whether major or minor, are dissonant intervals, whatever system of temperament be employed, as are also an indefinite number of other intervals not recognized in music.

Besides the difference as regards pleasing or unpleasing effect, it is to be remarked that consonant intervals can be identified by the ear with much greater accuracy than those which are dissonant. Musical instruments are generally tuned by octaves and fifths, because very slight errors of excess or defect in these intervals are easily detected by the ear. To tune a piano by the mere comparison of successive notes would be beyond the power of the most skilful musician. A sharply marked interval is always a consonant interval.

930. Jarring Effect of Dissonance.—According to the theory propounded by Helmholtz, the unpleasant effect of a dissonant interval consists essentially in the production of beats. These have a jarring effect upon the auditory apparatus, which becomes increasingly disagreeable as the beats increase in frequency up to a certain limit (about 33 per second for notes of medium pitch), and becomes gradually less disagreeable as the frequency is still further increased. The sensation produced by beats is comparable to that which the eye experiences from the *bobbing* of a gas flame in a room lighted by it; but the frequency which entails the maximum of annoyance is smaller for the eye than for the ear, on account of the greater persistence of visible impressions. The annoyance must evidently cease when the succession becomes so rapid as to produce the effect of a continuous impression.

We have already (§ 888) described a mode of producing beats with any degree of frequency at pleasure; and this experiment is one of the main foundations on which Helmholtz bases his view.

931. Beats of Harmonics.—The beats in the experiment above alluded to, are produced by the imperfect unison of two notes, and indicate the number of vibrations gained by one note upon the other. Their existence is easily and completely explained by the considerations adduced in § 888. But it is well known to musicians, and easily established by experiment, that beats are also produced between notes whose interval is approximately an octave, a fifth, or some other consonance; and that, in these cases also, the beats become more rapid as the interval becomes more faulty.

These beats are ascribed by Helmholtz to the common harmonic of the two fundamental notes. For example, in the case of the fifth (2 : 3), the third tone of the lower note would be identical with the second tone of the upper, if the interval were exact; and the beats which occur are due to the imperfect unison consequent on the deviation from exact truth. All beats are thus explained as due to imperfect *unison*.

This explanation is not merely conjectural, but is established by the following proofs:—

1. When an arrangement is employed by which the fifth is made false by a known amount, the number of beats is found to agree with the above explanation. Thus, if the interval is made to correspond to the ratio 200 : 301, it is observed that there are 2 beats to every 200 vibrations of the lower note. Now the harmonics which

are in approximate unison are represented by 600 and 602, and the difference of these is 2.

2. When the resonator corresponding to this common harmonic is held to the ear, it responds to the beats, showing that this harmonic is undergoing variations of strength; but when a resonator corresponding to either of the fundamental notes is employed, it does not respond to the beats, but indicates steady continuance of its appropriate note.

3. By a careful exercise of attention, a person with a good ear can hear, without any artificial aids, that it is the common harmonic which undergoes variations of intensity, and that the fundamental notes continue steady.

932. Beating Notes must be Near Together.—In order that two simple tones may yield audible beats, it is necessary that the musical interval between them should be small; in other words, that the ratio of their frequencies of vibration should be nearly equal to unity. Two simple notes of 300 and 320 vibrations per second will yield 20 beats in a second, and will be eminently discordant, the interval between them being only a semitone (15 : 16), but simple notes of 40 and 60 vibrations per second will not give beats, the interval between them being a fifth (2 : 3). The wider the interval between two simple notes, the feebler will be their beats; and accordingly, for a given frequency of beats, the harshness of the effect increases with the nearness of the notes to each other on the musical scale.[1] By taking joint account of the number of beats and the nearness of the beating tones, Helmholtz has endeavoured to express numerically the severity of the discords resulting from the combination of the note C of 256 vibrations per second with any possible note lying within an octave on the upper side of it, a particular constitution (approximately that of the violin) being assumed for both notes. He finds a complete absence of discord for the intervals of unison, the octave, and the fifth, and very small amounts of discord for the fourth, the sixth, and the third. By far the worst discords are found for the intervals of the semitone and major seventh,

[1] The explanation adopted by Helmholtz is, that a certain part of the ear—the *membrana basilaris*—is composed of tightly stretched elastic fibres, each of which is attuned to a particular simple tone, and is thrown into vibration when this tone, or one nearly in unison with it, is sounded. Two tones in approximate unison, when sounded together, affect several fibres in common, and cause them to beat. Tones not in approximate unison affect entirely distinct sets of fibres, and thus cannot produce interference.

and the next worst are for intervals a little greater or less than the fifth.

933. Imperfect Concord.—When there is a complete absence of discord between two notes, they are said to form a perfect concord. The intervals unison, fifth, octave, octave + fifth, and the interval from any note to any of its harmonics, are of this class. The third, fourth, and sixth are instances of imperfect concord. Suppose, for example, that the two notes sounded together are C of 256 and E of 320 vibrations per second, the interval between these notes being a true major third (4 : 5); and suppose each of these notes to consist of the first six simple tones.

The first six multiples of 4 are

4, 8, 12, 16, 20, 24.

The first six multiples of 5 are

5, 10, 15, 20, 25, 30.

In searching for elements of discord, we select (one from each line) two multiples differing by unity.

Those which satisfy this condition are

4 and 5; 16 and 15; 24 and 25.

But the first pair (4 and 5) may be neglected, because their ratio differs too much from unity. Discordance will result from each of the two remaining pairs; that is to say, the 4th element of the lower of our two given notes is in discordance with the 3d element of the upper; and the 6th element of the lower is in discordance with the 5th element of the higher. To find the frequencies of the beats, we must multiply all these numbers by 64, since 256 is 4 times 64, and 320 is 5 times 64. Instead of a difference of 1, we shall then find a difference of 64, that is to say, the number of beats per second is 64 in the case of each of the two discordant combinations which we have been considering.

934. Resultant Tones.—Under certain conditions it is found that two notes, when sounded together, produce by their combination other notes, which are not constituents of either. They are called *resultant tones*, and are of two kinds, *difference-tones* and *summation-tones*. A difference-tone has a frequency of vibration which is the difference of the frequencies of its components. A summation-tone has a frequency of vibration which is the sum of the

frequencies of its components. As the components may either be fundamental tones or overtones, two notes which are rich in harmonies may yield, by their combination, a large number of resultant tones.

The difference-tones were observed in the last century by Sorge and by Tartini, and were, until recently, attributed to beats. The frequency of beats is always the difference of the frequencies of vibration of the two elementary tones which produce them; and it was supposed that a rapid succession of beats produced a note of pitch corresponding to this frequency.

This explanation, if admitted, would furnish an exception to what otherwise appears to be the universal law, that every *elementary tone* arises from a corresponding *simple vibration*.[1] Such an exception should not be admitted without necessity; and in the present instance it is not only unnecessary, but also insufficient, inasmuch as it fails to render any account of the summation-tones.

Helmholtz has shown, by a mathematical investigation, that when two systems of simple waves agitate the same mass of air, their mutual influence must, according to the recognized laws of dynamics, give rise to two derived systems, having frequencies which are respectively the sum and the difference of the frequencies of the two primary systems. Both classes of resultant tones are thus completely accounted for.

The resultant tones—especially the summation-tones, which are fainter than the others—are only audible when the primary tones are loud; for their existence depends upon small quantities of the second order, the amplitudes of the primaries being regarded (in comparison with the wave-lengths) as small quantities of the first order.

If any further proof be required that the difference tones are not due to the coalescence of beats, it is furnished by the fact that, under favourable conditions, the rattle of the beats and the booming of the difference-tones can both be heard together.

R **935. Beats due to Resultant Tones.**—The existence of resultant tones serves to explain, in certain cases, the production of beats between notes which are wanting in harmonics. For example, if two *simple* sounds, of 100 and 201 vibrations per second respectively, are sounded together, one beat per second will be produced between

[1] The discovery of this law is due to Ohm.

40960 vibrations [illegible]

the difference-tone of 101 vibrations and the primary tone of 100 vibrations. By the beats to which they thus give rise, resultant tones exercise an influence on consonance and dissonance.

Resultant tones, when sufficiently loud, are themselves capable of performing the part of primaries, and yielding what are called *resultant tones of the second order*, by their combination with other primaries. Several higher orders of resultant tones can, under peculiarly favourable circumstances, be sometimes detected.

OPTICS.

CHAPTER LXVII.

PROPAGATION OF LIGHT.

936. Light.—Light is the immediate external cause of our visual impressions. Objects, except such as are styled *self-luminous*, become invisible when brought into a dark room. The presence of something additional is necessary to render them visible, and that mysterious agent, whatever its real nature may be, we call *light*.

Light, like sound, is believed to consist in vibration; but it does not, like sound, require the presence of air or other gross matter to enable its vibrations to be propagated from the source to the percipient. When we exhaust a receiver, objects in its interior do not become less visible; and the light of the heavenly bodies is not prevented from reaching us by the highly vacuous spaces which lie between.

It seems necessary to assume the existence of a medium far more subtle than ordinary matter; a medium which pervades alike the most vacuous spaces and the interior of all bodies, whether solid, liquid, or gaseous; and which is so highly elastic, in proportion to its density, that it is capable of transmitting vibrations with a velocity enormously transcending that of sound.

This hypothetical medium is called *æther*. From the extreme facility with which bodies move about in it, we might be disposed to call it a subtle *fluid;* but the undulations which it serves to propagate are not such as can be propagated by fluids. Its elastic properties are rather those of a solid; and its waves are analogous to the pulses which travel along the wires of a piano rather than to the waves of extension and compression by which sound is propagated through air. *Luminous vibrations are transverse, while those of sound are longitudinal.*

A self-luminous body, such as a red-hot poker or the flame of a

candle, is in a peculiar state of vibration. This vibration is communicated to the surrounding æther, and is thus propagated to the eye, enabling us to see the body. In the majority of cases, however, we see bodies not by their own but by reflected light; and we are enabled to recognize the various kinds of bodies by the different modifications which light undergoes in reflection from their surfaces.

As all bodies can become sonorous, so also all bodies can become self-luminous. To render them so, it is only necessary to raise them to a sufficiently high temperature, whether by the communication of heat from a furnace, or by the passage of an electric current, or by causing them to enter into chemical combination. It is to chemical combination, in the active form of combustion, that we are indebted for all the sources of artificial light in ordinary use.

The vibrations of the æther are capable of producing other effects besides illumination. They constitute what is called radiant heat, and they are also capable of producing chemical effects, as in photography. Vibrations of high frequency, or short period, are the most active chemically. Those of low frequency or long period have usually the most powerful heating effects; while those which affect the eye with the sense of light are of moderate frequency.

937. Rectilinear Propagation of Light.—All the remarks which have been made respecting the relations between period, frequency, and wave-length, in the case of sound, are equally applicable to light, inasmuch as all kinds of luminous waves (like all kinds of sonorous waves) have sensibly the same velocity in air; but this velocity is many hundreds of thousands of times greater for light than for sound, and the wave-lengths of light are at the same time very much shorter than those of sound. Frequency, being the quotient of velocity by wave-length, is accordingly about a million of millions of times greater for light than for sound. The colour of lowest pitch is deep red, its frequency being about 400 million million vibrations per second, and its wave-length in air 760 millionths of a millimetre. The colour of highest pitch is deep violet; its frequency is about 760 million million vibrations per second, and its wave-length in air 400 millionths of a millimetre. It thus appears that the range of seeing is much smaller than that of hearing, being only about one octave.

The excessive shortness of luminous as compared with sonorous waves is closely connected with the strength of the shadows cast by a light, as compared with the very moderate loss of intensity produced by interposing an obstacle in the case of sound. Sound may,

for ordinary purposes, be said to be capable of turning a corner, and light to be only capable of travelling in straight lines. The latter fact may be established by such an arrangement as is represented in Fig. 641. Two screens, each pierced with a hole, are arranged so that these holes are in a line with the flame of a candle. An eye placed in this line, behind the screens, is then able to see the flame; but a slight lateral displacement, either of the eye, the candle, or either of the screens, puts the flame out of sight. It is to be noted that, in this experiment, the same medium (air) extends from the eye to the candle. We shall hereafter find that, when light has to pass from one medium to another, it is often bent out of a straight line.

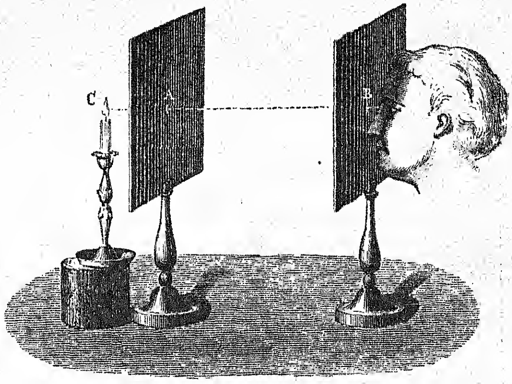


Fig. 641.—Rectilinear Propagation.

We have said that the strength of light-shadows as compared with sound-shadows is connected with the shortness of luminous waves. Theory shows that, if light is transmitted through a hole or slit, whose diameter is a very large multiple of the length of a light-wave, a strong shadow should be cast in all oblique directions; but that, if the hole or slit is so narrow that its diameter is comparable to the length of a wave, a large area not in the direct path of the beam will be illuminated. The experiment is easily performed in a dark room, by admitting sunlight through an exceedingly fine slit, and receiving it on a screen of white paper. The illuminated area will be marked with coloured bands, called diffraction-fringes; and if the slit is made narrower, these bands become wider.

On the other hand, Colladon, in his experiments on the transmission of sound through the water of the Lake of Geneva, established the presence of a very sharply defined sound-shadow in the water, behind the end of a projecting wall.

For the present we shall ignore diffraction,[1] and confine our atten-

[1] See Chap. lxxiv.

tion to the numerous phenomena which result from the rectilinear propagation of light.

938. Images produced by Small Apertures.—If a white screen is placed opposite a hole in the shutter of a room otherwise quite dark,

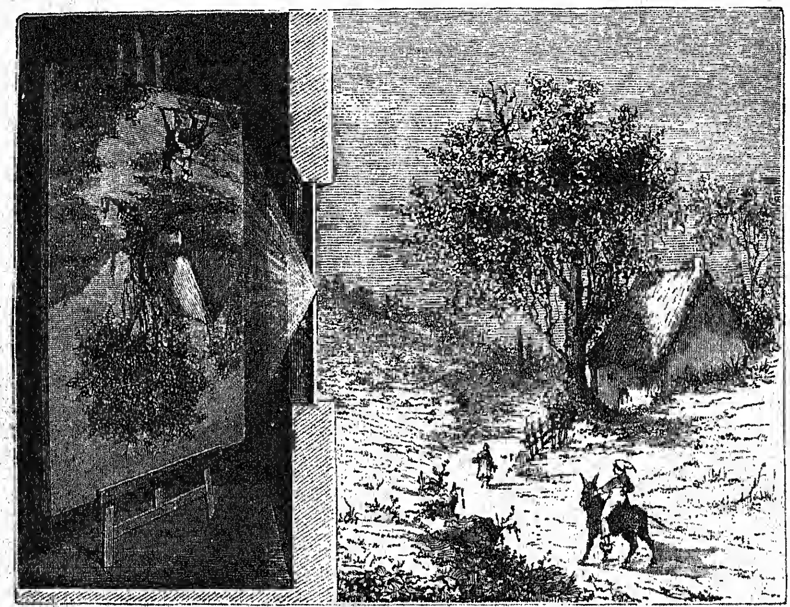


Fig. 642.—Image formed by Small Aperture.

an inverted picture of the external landscape will be formed upon it, in the natural colours. The outlines will be sharper in proportion as the hole is smaller, and distant objects will be more distinctly represented than those which are very near.

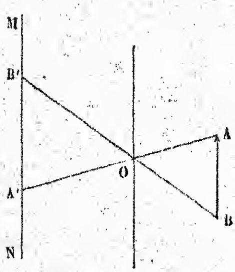


Fig. 643.—Explanation.

These results are easily explained. Consider, in fact, an external object A B (Fig. 643), and let O be the hole in the shutter. The point A sends rays in all directions into space, and among them a small pencil, which, after passing through the opening O, falls upon the screen at A′. A′ receives light from no other point but A, and A sends light to no part of the screen except A′. The colour and brightness of the spot A′ will accordingly depend upon the colour and brightness of A; in other words, A′ will be the

image of A. In like manner B′ will be the image of B, and points of the object between A and B will have their images between A′ and B′. An inverted image A′B′ will thus be formed of the object AB.

As the image thus formed of an external point is not a point, but a spot, whose size increases with that of the opening, there must always be a little blurring of the outlines from the overlapping of the spots which represent neighbouring points; but this will be comparatively slight if the opening is very small.

An experiment, substantially the same as the above, may be performed by piercing a card with a large pin-hole, and holding it between

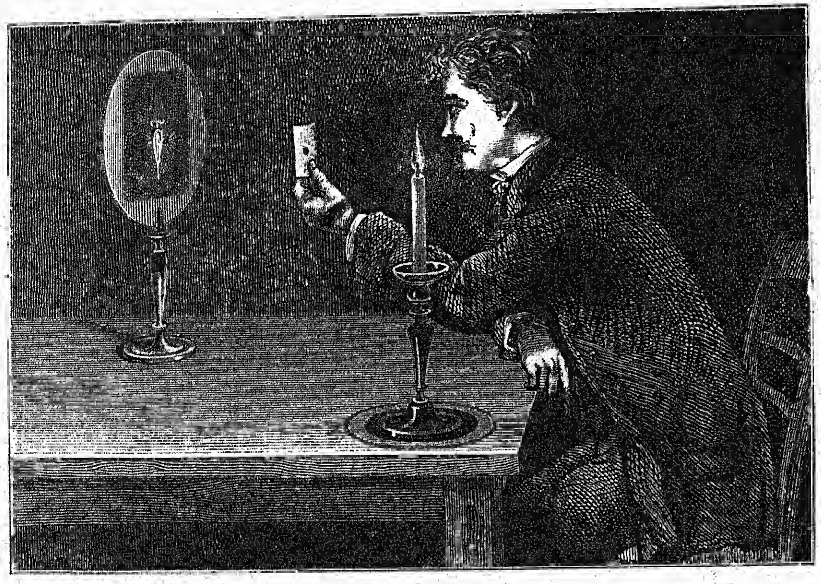


Fig. 644.—Image formed by Hole in a Card.

a candle and a screen, as in Fig. 644. An inverted image of the candle will thus be formed upon the screen.

When the sun shines through a small hole into a room with the blinds down (Fig. 645), the cone of rays thus admitted is easily traced by the lighting up of the particles of dust which lie in its course. The image of the sun which is formed at its further extremity will be either circular or elliptical, according as the incidence of the rays is normal or oblique. Fine images of the sun are sometimes thus formed by the chinks of a venetian-blind, especially when the sun is low, and there is a white wall opposite to receive the

image. In these circumstances it is sometimes possible to detect the presence of spots on the sun by examining the image.

When the sun's rays shine through the foliage of a tree (Fig. 646), the spots of light which they form upon the ground are always round or oval, whatever may be the shape of the interstices through which they have passed, provided always that these interstices are small. When the sun is undergoing eclipse, the progress of the eclipse can

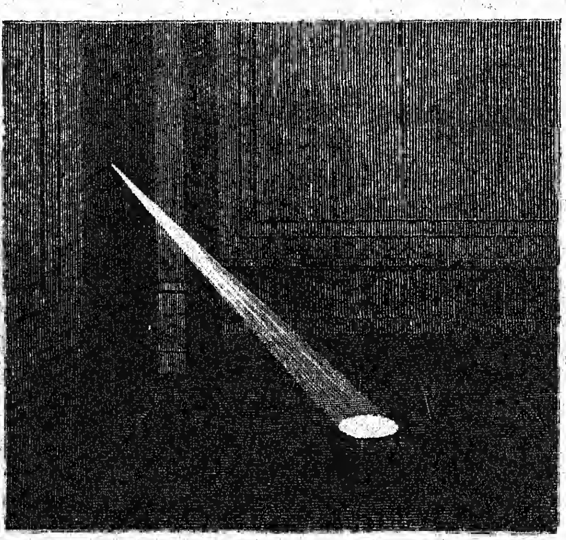


Fig. 645.—Conical Sunbeam.

be traced by watching the shape of these images, which resembles that of the uneclipsed portion of the sun's disc.

939. Theory of Shadows.—The rectilinear propagation of light is the foundation of the geometry of shadows. Let the source of light be a luminous point, and let an opaque body be placed so as to intercept a portion of its rays (Fig. 647). If we construct a conical surface touching the body all round, and having its vertex at the luminous point, it is evident that all the space within this surface on the further side of the opaque body is completely screened from the rays. The cone thus constructed is called the shadow-cone, and its intersection with any surface behind the opaque body defines the shadow cast upon that surface. In the case which we have been supposing—that of a luminous point—the shadow-cone and the shadow itself will be sharply defined.

Fig. 646.—Images of Sun formed by Foliage.

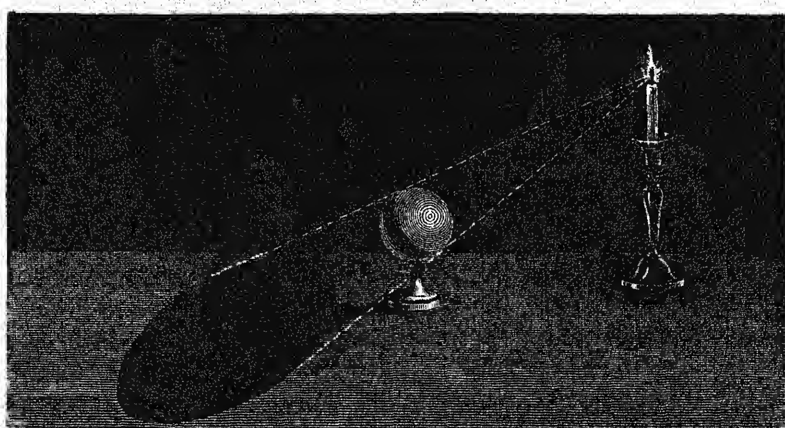


Fig. 647.—Shadow.

Actual sources of light, however, are not mere luminous points, but have finite dimensions. Hence some complication arises. Consider, in fact (Fig. 648), a luminous body situated between two opaque bodies, one of them larger, and the other smaller than itself. Conceive a cone touching the luminous body and either of the opaque bodies *externally*. This will be the cone of *total shadow*, or the cone of the *umbra*. All points lying within it are completely excluded from view of the luminous body. This cone narrows or enlarges as it recedes, according as the opaque body is smaller or larger than the luminous body. In the former case it terminates at a finite distance. In the latter case it extends to infinite distance.

Now conceive a double cone touching the luminous body and either of the opaque bodies *internally*. This cone will be wider than the cone of total shadow, and will include it. It is called the

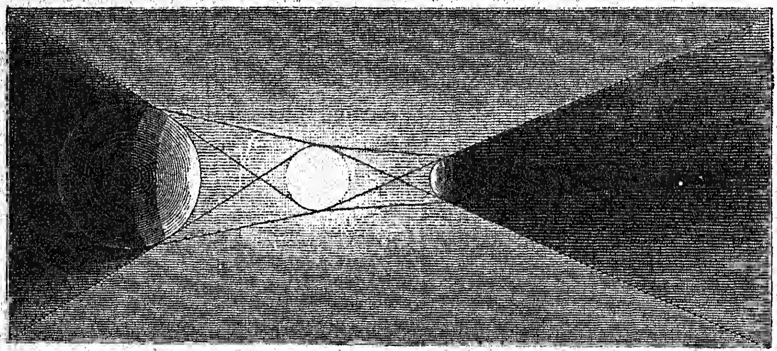


Fig. 648.—Umbra and Penumbra.

cone of *partial shadow*, or the cone of the *penumbra*. All points lying within it are excluded from the view of some portion of the luminous body, and are thus partially shaded by the opaque body. If they are near its outer boundary, they are very slightly shaded. If they are so far within it as to be near the total shadow, they are almost completely shaded. Accordingly, if the shadow of the opaque body is received upon a screen, it will not have sharply defined edges, but will show a gradual transition from the total shadow which covers a finite central area to a complete absence of shadow at the outer boundary of the penumbra. Thus neither the edges of the umbra nor those of the penumbra are sharply defined.

The umbra and penumbra show themselves on the surface of the

opaque body itself, the line of contact of the umbral cone being further back from the source of light than the line of contact of the penumbral cone. The zone between these two lines is in partial shadow, and separates the portion of the surface which is in total shadow from the part which is not shaded at all.

940. Velocity of Light.—Luminous undulations, unlike those of sound, advance with a velocity which may fairly be styled inconceivable, being about 300 million metres per second, or 186,000 miles per second. As the circumference of the earth is only 40 million metres, light would travel seven and a half times round the earth in a second.

Hopeless as it might appear to attempt the measurement of such an enormous velocity by mere terrestrial experiments, the feat has actually been performed, and that by two distinct methods. In Fizeau's experiments the distance between the two experimental stations was about $5\frac{1}{2}$ miles. In Foucault's experiments the whole apparatus was contained in one room, and the movement of light within this room served to determine the velocity.

We will first describe Fizeau's experiment.

941. Fizeau's Experiment.—Imagine a source of light placed directly in front of a plane mirror, at a great distance. The mirror will send back a reflected beam along the line of the incident beam, and an observer stationed behind the source will see its image in the mirror as a luminous point.

Now imagine a toothed-wheel, with its plane perpendicular to the path of the beam, revolving uniformly in front of the source, in such a position that its teeth pass directly between the source of light and the mirror. The incident beam will be stopped by the teeth, as they successively come up, but will pass through the spaces between them. Now the velocity of the wheel may be such that the light which has thus passed through a space shall be reflected back from the mirror just in time to meet a tooth and be stopped. In this case it will not reach the observer's eye, and the image may thus become permanently invisible to him. From the velocity of the wheel, and the number of its teeth, it will be possible to compute the time occupied by the light in travelling from the wheel to the mirror, and back again. If the velocity of the wheel is such that the light is sometimes intercepted on its return, and sometimes allowed to pass, the image will appear steadily visible, in consequence of the persistence of impressions on the retina, but with a

loss of brightness proportioned to the time that the light is intercepted. The wheel employed by Fizeau had 720 teeth, the distance between the two stations was 8663 metres, and 12·6 revolutions per second produced disappearance of the image. The width of the teeth being equal to the width of the spaces, the time required to turn through the width of a tooth was $\frac{1}{2} \times \frac{1}{720} \times \frac{1}{12 \cdot 6}$ of a second, that is $\frac{1}{18144}$ of a second.

In this time the light travelled a distance of $2 \times 8663 = 17326$ metres. The distance traversed by light in a second would therefore be $17{,}326 \times 18{,}144 = 314{,}262{,}944$ metres. This determination of M. Fizeau's is believed to be somewhat in excess of the truth.

A double velocity of the wheel would allow the reflected beam to pass through the space succeeding that through which the incident beam had passed; a triple velocity would again produce total eclipse, and so on. Several independent determinations of the velocity of light may thus be obtained.

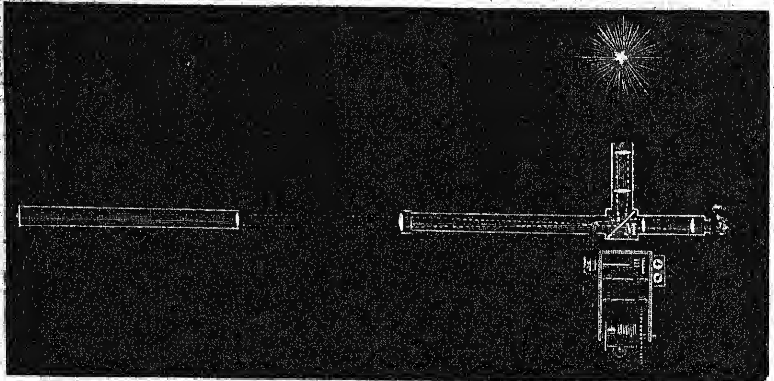


Fig. 649.—Fizeau's Experiment.

Thus far, we have merely indicated the principle of calculation. It will easily be understood that special means were necessary to prevent scattering of the light, and render the image visible at so great a distance. Fig. 649 will serve to give an idea of the apparatus actually employed.

A beam of light from a lamp, after passing through a lens, falls on a plate of unsilvered glass M, placed at an angle of 45°, by which it is reflected along the tube of a telescope; the object-glass of the telescope is so adjusted as to render the rays parallel on emergence, and in this condition they traverse the interval

between the two stations. At the second station they are collected by a lens, which brings them to a focus on the surface of a mirror, which sends them back along the same course by which they came. A portion of the light thus sent back to the glass plate M passes through it, and is viewed by the observer through an eye-piece.

The wheel R is driven by clock-work. Figs. 650, 651, 652 respec-

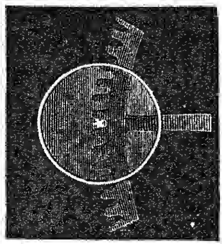


Fig. 650.—Wheel at Rest.

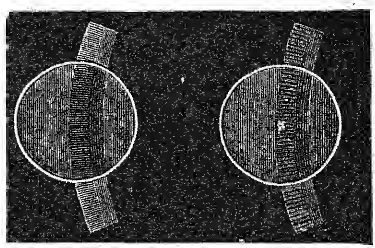


Fig. 651.—Total Eclipse. Fig. 652.—Partial Eclipse.

tively represent the appearance of the luminous point as seen between the teeth of the wheel when not revolving, the total eclipse produced by an appropriate speed of rotation, and the partial eclipse produced by a different speed.

More recently M. Cornu has carried out an extensive series of experiments on the same plan, with more powerful appliances, the distance between the two stations being 23 kilometres, and the extinctions being carried to the 21st order. His result is that the velocity of light (in millions of metres per second) is 300·33 in air, or 300·4 *in vacuo*.

942. Foucault's Experiment.—Foucault employed the principle of the rotating mirror, first adopted by Wheatstone in his experiments on the duration of the electric spark and the velocity of electricity (§ 591, 636). The following was the construction of his original apparatus:—

A beam of light enters a room by a square hole, which has a fine platinum wire stretched across it, to serve as a mark; it is then concentrated by an achromatic lens, and, before coming to a focus, falls upon a plane mirror, revolving about an axis in its own plane. In one part of the revolution the reflected beam is directed upon a concave mirror, whose centre of curvature is in the axis of rotation, so that the beam is reflected back to the revolving mirror, and

thence back to the hole at which it first entered. Before reaching the hole, it has to traverse a sheet of glass, placed at an angle of 45°, which reflects a portion of it towards the observer's eye; and the image which it forms (an image of the platinum wire) is viewed through a powerful eye-piece. The image is only formed during a small part of each revolution; but when 30 turns are made per second, the appearance presented, in consequence of the persistence of impressions, is that of a permanent image occupying a fixed position. When the speed is considerably greater, the mirror turns through a sensible angle while the light is travelling from it to the concave mirror and back again, and a sensible displacement of the image is accordingly observed. The actual speed of rotation was from 700 to 800 revolutions per second.[1]

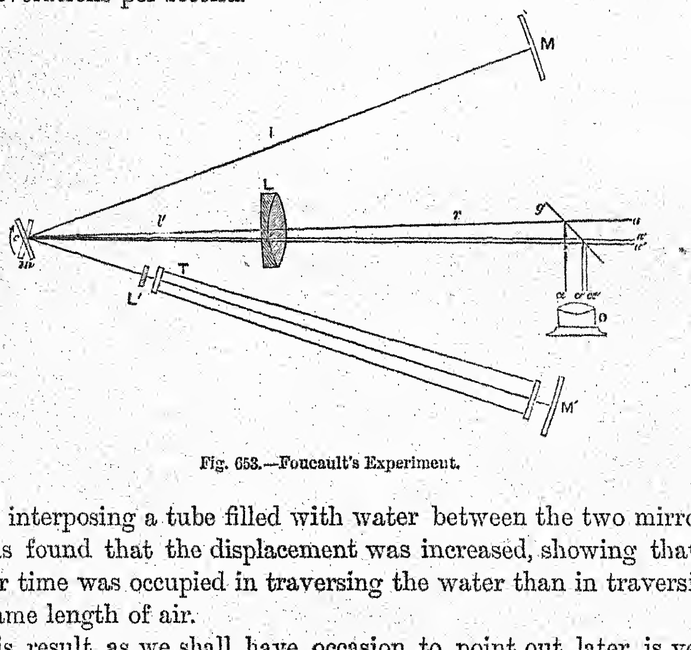


Fig. 653.—Foucault's Experiment.

On interposing a tube filled with water between the two mirrors, it was found that the displacement was increased, showing that a longer time was occupied in traversing the water than in traversing the same length of air.

This result, as we shall have occasion to point out later, is very important as confirming the undulatory theory and disproving the emission theory of light.

In Fig. 653, a is the position of the platinum wire, L is the achromatic lens, m the revolving mirror, c the axis of revolution, M

[1] It was found that, at this high speed, the amalgam at the back of ordinary looking-glasses was driven off by centrifugal force. The mirror actually employed was silvered in front with real silver.

the concave mirror, a' the image of the platinum wire, displaced from a in virtue of the rotation of the mirror; α, α' images of a, a', formed by the glass plate g, and viewed through the eye-piece O.

M′ is a second concave mirror, at the same distance as M from the revolving mirror; T is a tube filled with water, and having plane glass ends, and L′ a lens necessary for completing the focal adjustment; a'' and α'' are the images formed by the light which has traversed the water.[1]

Foucault's experiment, as thus described, was performed in 1850, very shortly after that of Fizeau. Some important improvements were afterwards introduced in the method, especially as regards the measurement of the speed of rotation of the mirror, which is evidently a principal element in the calculation. In the later arrangements the mirror was driven by means of a bellows, furnished with a special arrangement for keeping up a constant pressure of air, and driving a kind of siren, on which the mirror was mounted. Instead of making a separate determination of the speed of rotation in each experiment, means were employed for keeping it always at one constant value, namely, 400 revolutions per second. This was less than the speed attained in the earlier experiments; but, on the other hand, the length of the path traversed by the light between its two reflections from the revolving mirror was increased, by means of successive reflections, so as to be about 20 metres, instead of 4 as in the original experiments.

[1] The distances are such that La and Lc+cM are conjugate focal distances with respect to the lens L. An image of the wire a is thus formed at M, and an image of this image is formed at a, the mirror being supposed stationary; and this relation holds not only for the central point of the concave mirror, but for any part of it on which the light may happen to fall at the instant considered.

Let l denote the distance cM between the revolving and the fixed mirror, l' the distance cL of the revolving mirror from the centre of the lens, r the distance aL of the platinum wire from the centre of the lens, n the number of revolutions per second, V the space traversed by light in a second, t the time occupied by light in travelling from one mirror to the other and back, θ the angle turned by the mirror in this time, and δ the angle subtended at the centre of the lens by the distance $a\,a'$ between the wire and its displaced image.

Then obviously $t=\frac{2\,l}{V}$, but also $t=\frac{\theta}{2\,\pi\,n}$; hence $V=\frac{4\,\pi\,n\,l}{\theta}$.

Now the distance between the two images (corresponding to a, a' respectively) at the back of the revolving mirror is $(l+l')\,\delta$, and is also $2\,\theta\,l$ (§ 964). Hence $\theta=\frac{(l+l')\,\delta}{2\,l}$, and $V=\frac{8\,\pi\,n\,l^2}{(l+l')\delta}$. The observed distance $\alpha\,\alpha'$ between the two images is equal to the distance between a, a', that is to $r\,\delta$. Calling this distance d, we have finally,

$$V=\frac{8\,\pi\,n\,l^2\,r}{(l+l')\,d}.$$

The constant rate of revolution is maintained by comparison with a clock. A wheel with 400 teeth, driven by the clock, makes exactly one revolution per second. A tooth and a space alternately cover the part of the field where the image of the wire-grating (which has been substituted for the single wire) is formed. The same instantaneous flashes of light from the revolving mirror which form the image, also illuminate the rim of the wheel. If the wheel advances exactly one tooth and space between consecutive flashes, its illuminated positions are undistinguishable one from another, and the wheel accordingly appears stationary. When this is the case, it is known that the mirror is making exactly 400 turns per second. A slight departure from this rate either way, makes the wheel appear to be slowly revolving either forwards or backwards, and the bellows must be regulated until the stationary appearance is presented.

By means of this admirable combination, Foucault made by far the best determination that had up to that time been obtained of the velocity of light. His result was 298 million metres per second.

943. Michelson's Determination.—Captain Michelson of the United States' navy has recently effected some improvements in Foucault's method. He places the lens L not between the slit and the revolving mirror, but between the revolving and the fixed mirror, in such a position that the sum of the distances of the slit and lens from the revolving mirror is a very little greater than the focal length of the lens. The image of the slit is accordingly formed at a very great distance on the other side of the lens, and it is at this distance that the fixed mirror M is placed. The focal length of the lens was 150 ft. and the distance between the two mirrors nearly 2000 ft. The measured deviation of the image of the slit from the slit itself is due to the angle through which the mirror turns while light travels over twice this distance, or nearly 4000 ft., and the distance of the slit from the mirror being about 30 ft., the deviation of the image from the slit amounted to more than 133 millimetres, whereas the deviation obtained by Foucault was less than 1 millimetre.

The velocity deduced by Captain Michelson as the final result of his observations is 299·740 million metres per second in air, or 299·828 *in vacuo.*[1] This latter is about 186,300 miles per second.

944. Velocity of Light deduced from Observations of the Eclipses of Jupiter's Satellites.—The fact that light occupies a sensible time in travelling over celestial distances, was first established about 1675,

[1] For further details see *Nature*, Nov. 27 and Dec. 4, 1879.

by Roemer, a Danish astronomer, who also made the first computation of its velocity. He was led to this discovery by comparing the observed times of the eclipses of Jupiter's first satellite, as contained in records extending over many successive years.

The four satellites of Jupiter revolve nearly in the plane of the planet's orbit, and undergo very frequent eclipse by entering the cone of total shadow cast by Jupiter. The satellites and their eclipses are easily seen, even with telescopes of very moderate power; and being visible at the same absolute time at all parts of the earth's surface at which they are visible at all, they serve as signals for comparing local time at different places, and thus for determining longitudes. The first satellite (that is, the one nearest to Jupiter), from its more rapid motion and shorter time of revolution, affords both the best and the most frequent signals. The interval of time between two successive eclipses of this satellite is about 42½ hours, but was found by Roemer to vary by a regular law according to the position of the earth with respect to Jupiter. It is longest when the earth is increasing its distance from Jupiter most rapidly, and is shortest when the earth is diminishing its distance most rapidly. Starting from the time when the earth is nearest to Jupiter, as at T, J (Fig. 654), the intervals between successive eclipses are always longer than the mean value, until the greatest distance has been attained, as at T′ J′, and the sum of the excesses amounts to 16 min. 26·6 sec. From this time until the nearest distance is again attained, as at T″, J″, the intervals are always shorter than the mean, and the sum of the defects amounts to 16 min. 26·6 sec. It is evident, then, that the eclipses are visible 16 m. 26·6 s. earlier at the nearest than at the remotest point of the earth's orbit; in other words, that this is the time required for the propagation of light across the diameter of the orbit. Taking this diameter as 184 millions of miles,[1] we have a resulting velocity of about 186,500 miles per second.

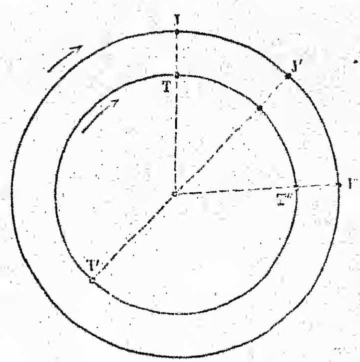


Fig. 654.—Earth and Jupiter.

945. Velocity of Light deduced from Aberration.—About fifty

[1] The sun's mean distance from the earth was, until recently, estimated at 95 millions of miles. It is now estimated at 92 or 92½ millions.

years after Roemer's discovery, Bradley, the English astronomer, employed the velocity of light to explain the astronomical phenomenon called *aberration.* This consists in a regular periodic displacement of the stars as seen from the earth, the period of the displacement being a year. If the direction in which the earth is moving in its orbit at any instant be regarded as the *forward* direction every star constantly appears on the forward side of its true place, so that, as the earth moves once round its orbit in a year, each star describes in this time a small apparent orbit about its true place.

The phenomenon is explained in the same way as the familiar fact, that a shower of rain falling vertically, seems, to a person running forwards, to be coming in his face. The relative motion of the rain-drops with respect to his body, is found by compounding the actual velocity of the drops (whether vertical or oblique) with a velocity equal and opposite to that with which he runs. Thus if A B (Fig. 655) represents the velocity with which he runs, and C A, the true velocity of the drops, the apparent velocity of the drops will be represented by D A. If a tube pointed along A D moves forward parallel to itself with the velocity A B, a drop entering at its upper end will pass through its whole length without wetting its sides; for while the drop is falling along D B (we suppose with uniform velocity) the tube moves along A B, so that the lower end of the tube reaches B at the same time as the rain-drop.

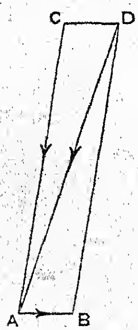


Fig. 655. Aberration.

In like manner, if A B is the velocity of the earth, and C A the velocity of light, a telescope must be pointed along A D to see a star which really lies in the direction of A C or B D produced. When the angle B A C is a right angle (in other words, when the star lies in a direction perpendicular to that in which the earth is moving), the angle C A D, which is called the aberration of the star, is 20″·5, and the tangent of this angle is the ratio of the velocity of the earth to the velocity of light. Hence it is found by computation that the velocity of light is about ten thousand times greater than that with which the earth moves in its orbit. The latter is easily computed, if the sun's distance is known, and is about 18½ miles per second. Hence the velocity of light is about 185,000 miles per second. It will be noted that both these astronomical methods of computing the velocity of light, depend upon the knowledge of the sun's distance

from the earth, and that, if this distance is overestimated, the computed velocity of light will be too great in the same ratio.

Conversely, the velocity of light, as determined by Foucault's method, can be employed in connection either with aberration or the eclipses of the satellites, for computing the sun's distance; and the first correct determination of the sun's distance was, in fact, that deduced by Foucault from his own results.

946. Photometry.—Photometry is the measurement of the relative amounts of light emitted by different sources. The methods employed for this purpose all consist in determinations of the relative distances at which two sources produce equal intensities of illumination. The eye would be quite incompetent to measure the ratio of two unequal illuminations; but a pretty accurate judgment can be formed as to equality or inequality of illumination, at least when the surfaces compared are similar, and the lights by which they are illuminated are of the same colour. The law of inverse squares is always made the basis of the resulting calculations; and this law may itself be verified by showing that the illumination produced by one candle at a given distance is equal to that produced by four candles at a distance twice as great.

947. Bouguer's Photometer.—Bouguer's photometer consists of a

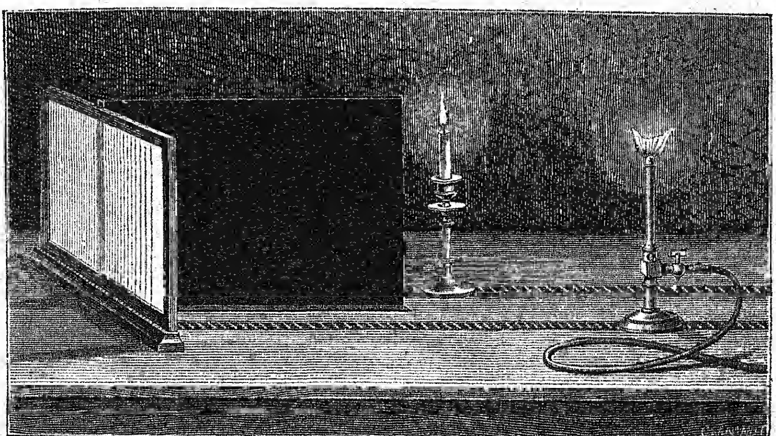


Fig. 656.—Bouguer's Photometer.

semi-transparent screen, of white tissue paper, ground glass, or thin white porcelain, divided into two parts by an opaque partition at right angles to it. The two lamps which are to be compared are

placed one on each side of this partition, so that each of them illuminates one-half of the transparent screen. The distances of the two lamps are adjusted until the two portions of the screen, as seen from the back, appear equally bright. The distances are then measured, and their squares are assumed to be directly proportional to the illuminating powers of the lamps.

948. Rumford's Photometer.—Rumford's photometer is based on the comparison of shadows. A cylindric rod is so placed that each of the two lamps casts a shadow of it on a screen; and the distances

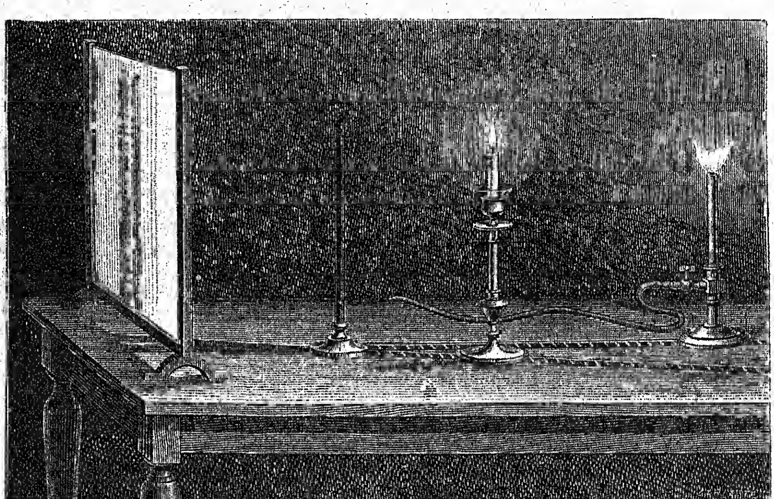


Fig. 657.—Rumford's Photometer.

are adjusted until the two shadows are equally dark. As the shadow thrown by one lamp is illuminated by the other lamp, the comparison of shadows is really a comparison of illuminations.

949. Foucault's Photometer.—The two photometers just described are alike in principle. In each of them the two surfaces compared are illuminated each by one only of the sources of light. In Rumford's the remainder of the screen is illuminated by both. In Bouguer's it consists merely of an intervening strip which is illuminated by neither. If the partition is movable, the effect of moving it further from the screen will be to make this dark strip narrower until it disappears altogether; and if it be advanced still further, the two illuminated portions will overlap. In Foucault's photometer there is an adjusting screw, for the purpose of advancing the parti-

tion so far that the dark strip shall just vanish. The two illuminated portions, being then exactly contiguous, can be more easily and certainly compared.

950. Bunsen's Photometer.—In the instruments above described the two sources to be compared are both on the same side of the screen, and illuminate different portions of it. Bunsen introduced the plan of placing the sources on opposite sides of the screen, and making the screen consist of two parts, one of them more translucent than the other. In his original pattern the screen was a sheet of white paper, with a large grease spot in the centre. In Dr. Letheby's pattern it is composed of three sheets of paper, laid face to face, the middle one being very thin, and the other two being cut away in the centre, so that the central part of the screen consists of one thickness, and the outer part of three.

When such a screen is more strongly illuminated on one side than on the other, the more translucent part appears brighter than the less translucent when seen from the darker side, while the reverse appearance is presented on the brighter side. It is therefore the business of the observer so to adjust the distances that the central and circumferential parts appear equally bright. When they appear equally bright from one side they will also appear equally bright from the other; but as there is always some little difference of tint, the observer's judgment is aided by seeing both sides at once. This is accomplished in Dr. Letheby's photometer by means of two mirrors, one for each eye, as represented in the accompanying ground-plan (Fig. 658).

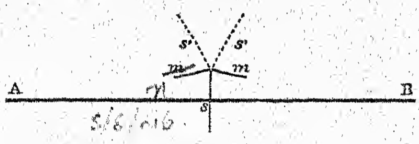


Fig. 658.—Letheby's Photometer.

s is the screen, and $n\,m$ are the two mirrors, in which images $s'\,s'$ are seen by an observer in front. The frame which carries the screen and mirrors travels along a graduated bar A B, on which the distances of the screen from the two lights are indicated.

In all delicate photometric observations, the eye should be shielded from direct view of the lights, and, as much as possible, from all extraneous lights. The objects to be compared should be brighter than anything else in the field of view.

951. Photometers for very Powerful Lights.—In comparing two very unequal lights, for example, a powerful electric light and a

standard candle, it is scarcely possible to obtain an observing-room long enough for a direct comparison by any of the above methods. To overcome this difficulty a lens (either convex or concave), of short focal length, may be employed to form an image of the more powerful source near its principal focus. Then all the light which this source sends to the lens may be regarded as diverging from the image and filling a solid angle equal to that which the lens subtends at the image. In other words, the illuminations of the lens itself due to the source and the image are equal. Hence, if S and I are the distances of the source and image from the lens, the image is weaker than the source in the ratio of I^2 to S^2, and a direct comparison can be made between the light from the image and that from a standard candle. Thus, if a screen at distance D from the image has the same illumination from the image as from a candle at distance C on the other side, the image is equal to $\frac{D^2}{C^2}$ candles, and the source itself to $\frac{D^2}{C^2} \frac{S^2}{I^2}$ candles. A correction must, however, be applied to this result for the light lost by reflection at the surfaces of the lens.

CHAPTER LXVIII.

REFLECTION OF LIGHT.

952. Reflection.—If a beam of the sun's rays A B (Fig. 659) be admitted through a small hole in the shutter of a dark room, and allowed to fall on a polished plane surface, it will be seen to continue its course in a different direction B C. This is an example of reflection.

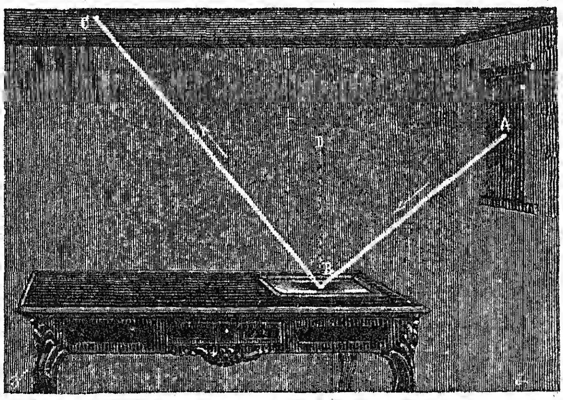


Fig. 659.—Reflection of Light.

A B is called the incident beam, and B C the reflected beam. The angle A B D contained between an incident ray and the normal is called the angle of incidence; and the angle C B D contained between the corresponding reflected ray and the normal is called the angle of reflection. The plane A B D containing the incident ray and the normal is called the plane of incidence.

953. Laws of Reflection.—The reflection of light from polished surfaces takes place according to the following laws:—

1. The reflected ray lies in the plane of incidence.

2. The angle of reflection is equal to the angle of incidence.

These laws may be verified by means of the apparatus represented in Fig. 660. A vertical divided circle has a small polished plate fixed at its centre, at right angles to its plane, and two tubes travelling on its circumference with their axes always directed towards the centre. The zero of the divisions is the highest point of the circle, the plate being horizontal.

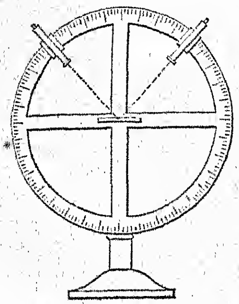


Fig. 660.—Verification of Laws of Reflection.

A source of light, such as the flame of a candle, is placed so that its rays shine through one of the tubes upon the plate at the centre. As the tubes are blackened internally, no light passes through except in a direction almost precisely parallel to the axis of the tube. The observer then looks through the other tube, and moves it along the circumference till he finds the position in which the reflected light is visible through it. On examining the graduations, it will be found that the two tubes are at the same distance from the zero point, on opposite sides. Hence the angles of incidence and reflection are equal. Moreover the plane of the circle is the plane of incidence, and this also contains the reflected rays. Both the laws are thus verified.

954. **Artificial Horizon.**—These laws furnish the basis of a method of observation which is frequently employed for determining the altitude of a star, and which, by the consistency of its results, furnishes a very rigorous proof of the laws.

A vertical divided circle (Fig. 661) is set in a vertical plane by proper adjustments. A telescope movable about the axis of the circle is pointed to a particular star, so that its line of collimation I′S′ passes through the apparent place of the star. Another telescope,[1] similarly mounted on the other side of the circle, is directed downwards along the line I′R towards the image of the star as seen in a trough of mercury I. Assuming the truth of the laws of reflection as above stated, the altitude of the star is half the angle between the directions of the two telescopes; for the ray SI from the star to the mercury is parallel to the line S′I′, by reason of the excessively great distance of the star; and since the rays SI, IR are equally inclined to the normal IN, which is a vertical line, the lines I′S, I′R are also equally inclined to the vertical, or, what is the same thing,

[1] In practice, a single telescope usually serves for both observations.

are equally inclined to a horizontal plane. A reflecting surface of mercury thus used is called a mercury horizon, or an *artificial*

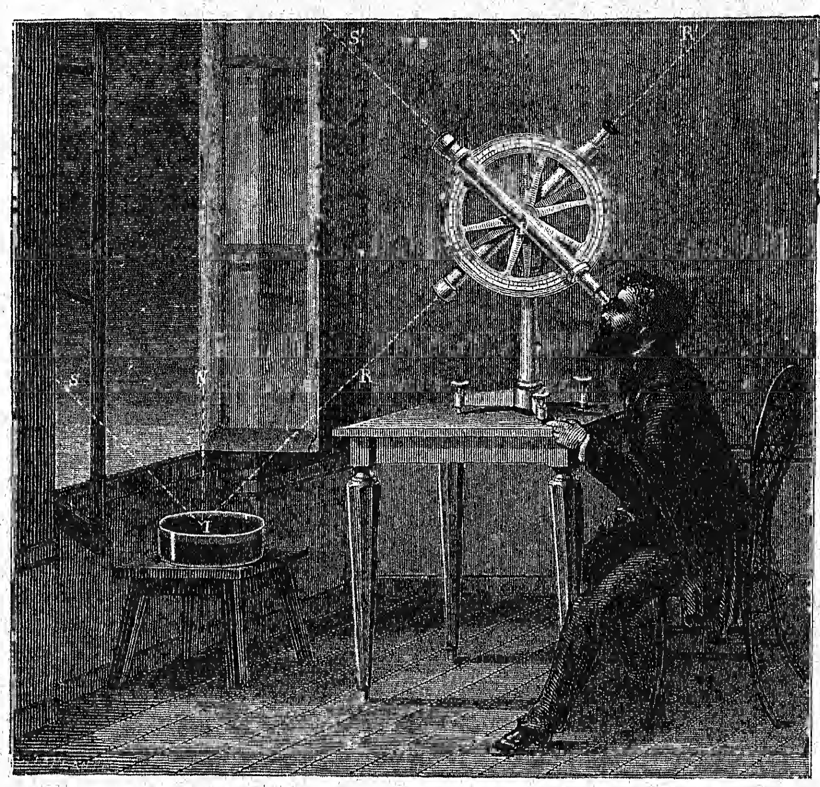


Fig. 661.—Artificial Horizon.

horizon. Observations thus made give even more accurate results than those in which the natural horizon presented by the sea is made the standard of reference.

955. **Irregular Reflection.**—The reflection which we have thus far been discussing is called *regular reflection.* It is more marked as the reflecting surface is more highly polished, and (except in the case of metals) as the incidence is more oblique. But there is another kind of reflection, in virtue of which bodies, when illuminated, send out light in all directions, and thus become visible. This is called *irregular reflection* or *diffusion.* Regular reflection does not render the reflecting body visible, but exhibits images of surrounding objects. A perfectly reflecting mirror would be itself unseen, and

2. The angle of reflection is equal to the angle of incidence.

These laws may be verified by means of the apparatus represented in Fig. 660. A vertical divided circle has a small polished plate fixed at its centre, at right angles to its plane, and two tubes travelling on its circumference with their axes always directed towards the centre. The zero of the divisions is the highest point of the circle, the plate being horizontal.

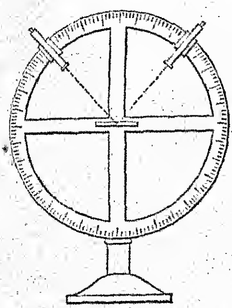


Fig. 660.—Verification of Laws of Reflection.

A source of light, such as the flame of a candle, is placed so that its rays shine through one of the tubes upon the plate at the centre. As the tubes are blackened internally, no light passes through except in a direction almost precisely parallel to the axis of the tube. The observer then looks through the other tube, and moves it along the circumference till he finds the position in which the reflected light is visible through it. On examining the graduations, it will be found that the two tubes are at the same distance from the zero point, on opposite sides. Hence the angles of incidence and reflection are equal. Moreover the plane of the circle is the plane of incidence, and this also contains the reflected rays. Both the laws are thus verified.

954. **Artificial Horizon.**—These laws furnish the basis of a method of observation which is frequently employed for determining the altitude of a star, and which, by the consistency of its results, furnishes a very rigorous proof of the laws.

A vertical divided circle (Fig. 661) is set in a vertical plane by proper adjustments. A telescope movable about the axis of the circle is pointed to a particular star, so that its line of collimation I′S′ passes through the apparent place of the star. Another telescope,[1] similarly mounted on the other side of the circle, is directed downwards along the line I′R towards the image of the star as seen in a trough of mercury I. Assuming the truth of the laws of reflection as above stated, the altitude of the star is half the angle between the directions of the two telescopes; for the ray S I from the star to the mercury is parallel to the line S′I′, by reason of the excessively great distance of the star; and since the rays S I, I R are equally inclined to the normal IN, which is a vertical line, the lines I′S, I′R are also equally inclined to the vertical, or, what is the same thing,

[1] In practice, a single telescope usually serves for both observations.

are equally inclined to a horizontal plane. A reflecting surface of mercury thus used is called a mercury horizon, or an *artificial*

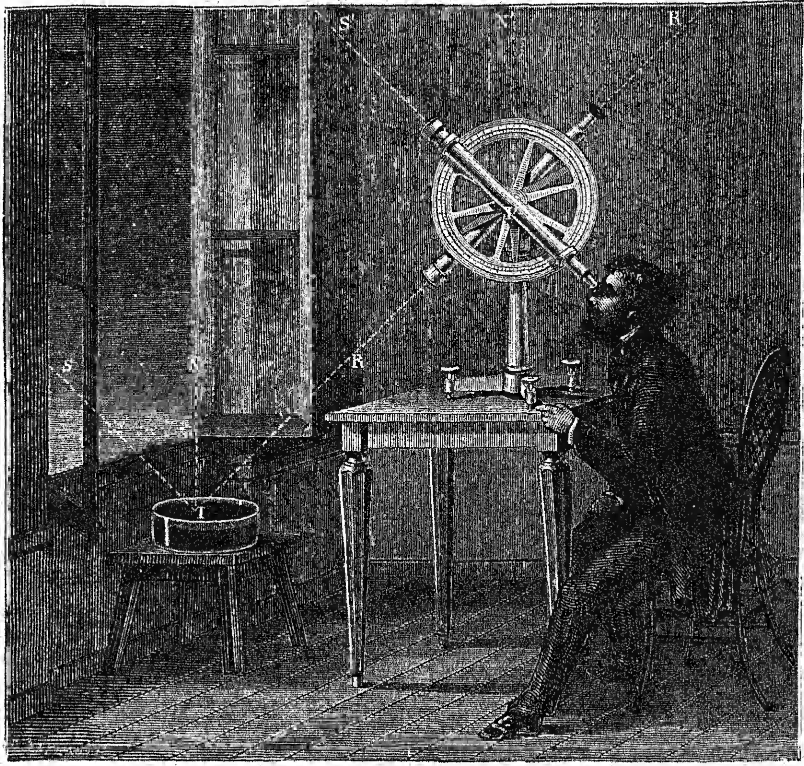


Fig. 661.—Artificial Horizon.

horizon. Observations thus made give even more accurate results than those in which the natural horizon presented by the sea is made the standard of reference.

955. Irregular Reflection.—The reflection which we have thus far been discussing is called *regular reflection.* It is more marked as the reflecting surface is more highly polished, and (except in the case of metals) as the incidence is more oblique. But there is another kind of reflection, in virtue of which bodies, when illuminated, send out light in all directions, and thus become visible. This is called *irregular reflection* or *diffusion.* Regular reflection does not render the reflecting body visible, but exhibits images of surrounding objects. A perfectly reflecting mirror would be itself unseen, and

actual mirrors are only visible in virtue of the small quantity of diffused light which they usually emit. The transformation of incident into diffused light is usually selective; so that, though the incident beam may be white, the diffused light is usually coloured. The power which a body possesses of making such selection constitutes its colour.

The word *reflection* is often used by itself to denote what we have here called *regular reflection*, and we shall generally so employ it.

956. Mirrors.—The mirrors of the ancients were of metal, usually of the compound now known as speculum-metal. Looking-*glasses* date from the twelfth century. They are plates of glass, coated at the back with an amalgam of quicksilver and tin, which forms the reflecting surface. This arrangement has the great advantage of excluding the air, and thus preventing oxidation. It is attended, however, with the disadvantage that the surface of the glass and the surface of the amalgam form two mirrors; and the superposition of the two sets of images produces a confusion which would be intolerable in delicate optical arrangements. The mirrors, or *specula* as they are called, of reflecting telescopes are usually made of *speculum-metal,* which is a bronze composed of about 32 parts of copper to 15 of tin. Lead, antimony, and arsenic are sometimes added. Of late years specula of glass coated in *front* with real silver have been extensively used; they are known as *silvered specula.* A coating of platinum has also been tried, but not with much success. The mirrors employed in optics are usually either *plane* or *spherical.*

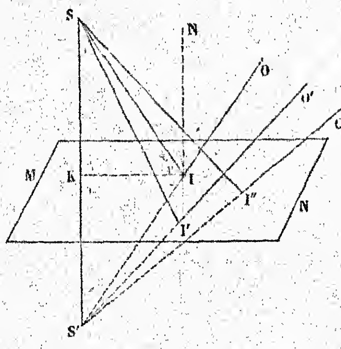


Fig. 662.—Plane Mirror.

957. Plane Mirrors.—By a plane mirror we mean any plane reflecting surface. Its effect, as is well known, is to produce, behind the mirror, images exactly similar, both in form and size, to the real objects in front of it. This phenomenon is easily explained by the laws of reflection.

Let M N (Fig. 662) be a plane mirror, and S a luminous point. Rays SI, SI′, SI″ proceeding from this point give rise to reflected rays I O, I′ O′, I″ O″; and each of these, if produced backwards, will meet the normal S K in a point S′, which is at the same distance behind the mirror that S is in front of

it.[1] The reflected rays have therefore the same directions as if they had come from S′, and the eye receives the same impression as if S′ were a luminous point.

Fig. 663 represents a pencil of rays emitted by the highest point of a candle-flame, and reflected from a plane mirror to the eye of an observer. The reflected rays are divergent (like the incident rays), and if produced backwards would meet in a point, which is the position of the image of the top of the flame.

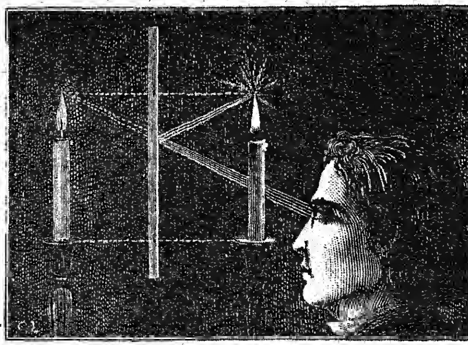


Fig. 663.—Image of Candle.

As an object is made up of points, these principles show that the image of an object formed by a plane mirror must be equal to the object, and symmetrically situated with respect to the plane of the mirror. For example, if AB (Fig. 664) is an object in front of the mirror, an eye placed at O will see the image of the point A at A′, the image of B at B′, and so on for all the other points of the object. The position of the image A′B′ depends only on the positions of the object and of the mirror, and remains stationary as the eye is moved about. It is possible, however, to find positions from which the eye will not see the image at all, the conditions of visibility being the same as if the image were a real object, and the mirror were an opening through which it could be seen.

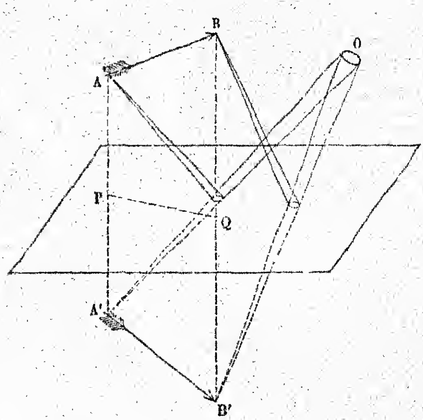


Fig. 664.—Incident and Reflected Pencils.

The images formed by a plane mirror are *erect*. They are not however exact duplicates of the objects from which they are formed,

[1] This is evident from the comparison of the two triangles S K I, S′K I, bearing in mind that the angle N I S is equal to the alternate angle I S K, and N I O to K S′ I.

but differ from them precisely in the same way as the left foot or hand differs from the right. The image of a printed page is like the

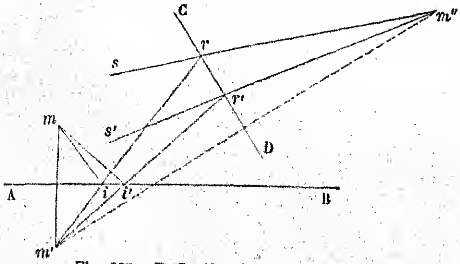


Fig. 665.—Reflection from two Mirrors.

appearance of the page as seen through the paper from the back, or like the type from which the page was printed.

958. **Images of Images.**—When rays from a luminous point m have been reflected from a mirror A B (Fig. 665), their subsequent course is the same as if they had come from the image m' at the back of the mirror. Hence, if they fall upon a second mirror C D, an image m'' of the first image will be formed at the back of the second mirror. If, after this, they undergo a third reflection, an image of m'' will be formed, and so on indefinitely. The figure shows the actual paths of two rays $m\,i\,r\,s$, $m\,i'\,r'\,s'$. They diverge first from

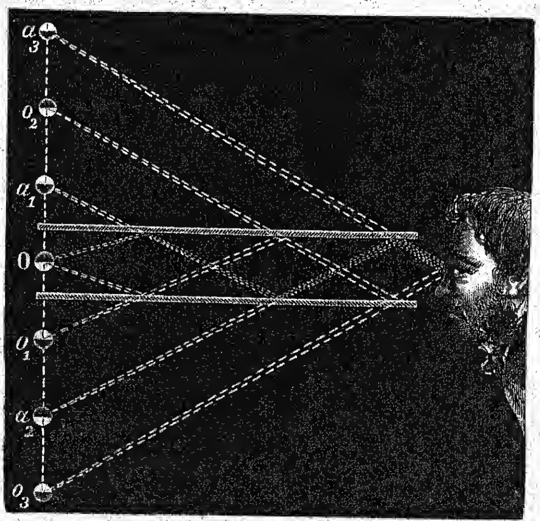


Fig. 666.—Parallel Mirrors.

m, then from m', and lastly from m''. This is the principle of the multiple images formed by two or more mirrors, as in the following experiments.

959. **Parallel Mirrors.**—Let an object O be placed between two

parallel mirrors which face each other, as in Fig 666. The first reflections will form images $a_1\,o_1$. The second reflections will form images $a_2\,o_2$ of the first images; and the third reflections will form images $a_3\,o_3$ of the second images. The figure represents an eye receiving *the rays which form the third images*, and shows the paths which *these rays* have taken in their whole course from the object O to the eye. The rays by which the same eye sees the other images are omitted, to avoid confusing the figure. A long row of images can thus be seen at once, becoming more and more dim as they recede in the distance, inasmuch as each reflection involves a loss of light.

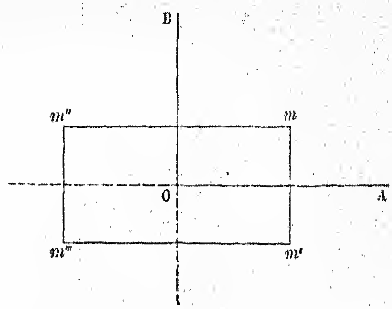


Fig. 667.—Mirrors at Right Angles.

If the mirrors are truly parallel, all the images will be ranged in one straight line, which will be normal to the mirrors. If the mirrors are inclined at any angle, the images will be ranged on the circumference of a circle, whose centre

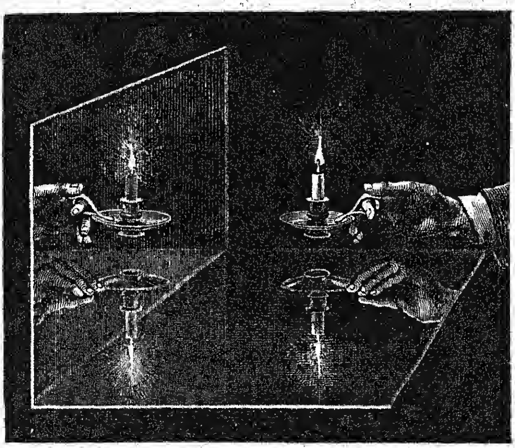


Fig. 668.—Mirrors at Right Angles.

is on the line in which the reflecting surfaces would intersect if produced. This principle is sometimes employed as a means of adjusting mirrors to exact parallelism.

960. Mirrors at Right Angles.—Let two mirrors O A, O B (Fig. 667),

be set at right angles to each other, facing inwards, and let m be a luminous point placed between them. Images m' m'' will be formed by first reflections, and two coincident images will be formed at m''' by second reflections. No third reflection will occur, for the point m''', being behind the planes of both the mirrors, cannot be reflected in either of them. Counting the two coincident images as one, and also counting the object as one, there will be in all four images, placed at the four corners of a rectangle. Fig. 668 will give an idea of the appearance actually presented when one of the mirrors is vertical and the other horizontal. When both the mirrors are vertical, an observer sees his own image constantly bisected by their common section in a way which appears at first sight very paradoxical.

961. Mirrors Inclined at 60 Degrees.—A symmetrical distribution of images may be obtained by placing a pair of mirrors at any angle which is an aliquot part of 360°. If, for example, they be inclined at 60° to each other, the number of images, counting the object itself as one, will be six. Their position is illustrated by Fig. 669. The object is placed in the sector A C B. The images formed by first reflections are situated in the two neighbouring sectors B C A′, A C B′; the images formed by second reflections are in the sectors B′ C A″, A′ C B″, and these yield, by third reflections, two coincident images in the sector B″ C A″, which is vertically opposite to the sector A C B in which the object lies, and is therefore behind the planes of both mirrors, so that no further reflection can occur.

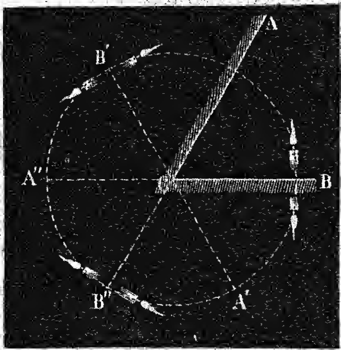


Fig. 669.—Images in Kaleidoscope.

962. Kaleidoscope.—The symmetrical distribution of images, obtained by two mirrors inclined at an angle which is an aliquot part of four right angles, is the principle of the *kaleidoscope*, an optical toy invented by Sir David Brewster. It consists of a tube containing two glass plates, extending along its whole length, and inclined at an angle of 60°. One end of the tube is closed by a metal plate, with the exception of a hole in the centre, through which the observer looks in; at the other end there are two plates, one of ground and the other of clear glass (the latter being next the eye), with a number of little pieces of coloured glass lying loosely between them. These

coloured objects, together with their images in the mirrors, form symmetrical patterns of great beauty, which can be varied by turning or shaking the tube, so as to cause the pieces of glass to change their positions.

A third reflecting plate is sometimes employed, the cross-section of the three forming an equilateral triangle. As each pair of plates produces a kaleidoscopic pattern, the arrangement is nearly equivalent to a combination of three kaleidoscopes.

The kaleidoscope is capable of rendering important aid to designers.

Fig. 670.—Kaleidoscopic Pattern.

Fig. 670 represents a pattern produced by the equilateral arrangement of three reflectors just described. Omit

963. Pepper's Ghost.—Many ingenious illusions have been contrived, depending on the laws of reflection from plane surfaces. We shall mention two of the most modern.

In the *magic cabinet*, there are two vertical mirrors hinged at the two back corners of the cabinet, and meeting each other at a right angle, so as to make angles of 45° with the sides, and also with the back. A spectator seeing the images of the two sides, mistakes them for the back, which they precisely resemble; and performers may be concealed behind the mirrors when the cabinet appears empty. If one of the persons thus concealed raises his head above the mirrors, it will appear to be suspended in mid-air without a body.

The striking spectral illusion known as *Pepper's Ghost* is produced by reflection from a large sheet of unsilvered glass, which is so arranged that the actors on the stage are seen through it, while other actors, placed in strong illumination, and out of the direct view of the spectators, are seen by reflection in it, and appear as ghosts on the stage.

964. Deviation produced by Rotation of Mirror.—Let A B (Fig. 671) represent a mirror perpendicular to the plane of the paper, and capable of being rotated about an axis through C, also perpendicular to the paper; and let I C represent an incident ray. When the mirror is in the position A B, perpendicular to I C, the ray will be reflected directly back upon its course; but when the mirror is turned through the acute angle A C A′, the reflected ray will take the direction C R, making with the normal C N an angle N C R, equal to the angle of incidence N C I. The deviation I C R of the reflected ray, produced by rotating the mirror, is therefore double of the angle I C N or A C A′, through which the mirror has been turned; and if, starting from the position A′B′, we turn the mirror through a further angle θ, the reflected ray C R will be turned through a further angle 2θ. It thus appears, that, *when a plane mirror is rotated in the plane of incidence, the direction of the reflected ray is changed by double the angle through which the mirror is turned.* Conversely, if we assign a constant direction C I to the reflected ray, the direction of the incident ray R C must vary by double the angle through which the mirror is turned.

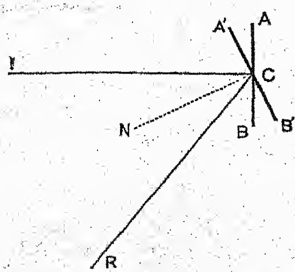


Fig. 671.—Effect of rotating a Mirror.

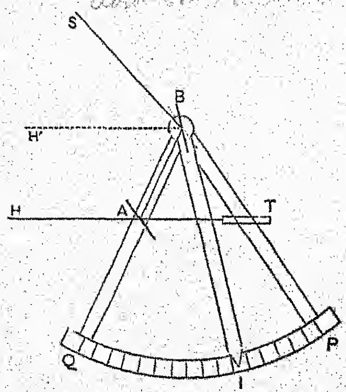


Fig. 672.—Sextant.

965. Hadley's Sextant.—The above principle is illustrated in the nautical instrument called the *sextant* or *quadrant*, which was invented by Newton, and reinvented by Hadley. It serves for measuring the angle between any two distant objects as seen from the station occupied by the observer. Its essential parts are represented in Fig. 672.

It has two plane mirrors A, B, one of which, A, is fixed to the frame of the instrument, and is only partially silvered, so that a distant object in the direction A H can be seen through the unsilvered part. The other mirror B is mounted on a movable arm B I, which carries an index I, traversing a graduated arc P Q. When the two mirrors are parallel, the index is at P, the zero of the graduations, and a ray H′ B incident on B parallel to H A, will be reflected first along B A, and then along A T, the continuation of H A. The observer looking through the telescope T thus sees, by two reflections, the same objects which he also sees directly through the unsilvered part of the mirror. Now let the index be advanced through an angle θ; then, by the principles of last section, the incident ray S B makes with H′ B, or H A, an angle 2θ. The angle between S B and H A would therefore be given by reading off the angle through which the index has been advanced, and doubling; but in practice the arc P Q is always graduated on the principle of marking half degrees as whole ones, so that the reading at I is the required angle 2θ. In using the instrument, the two objects which are to be observed are brought into apparent coincidence, one of them being seen directly, and the other by successive reflection from the two mirrors. This coincidence is not disturbed by the motion of the ship; but unpractised observers often find a difficulty in keeping both objects in the field of view. Dark glasses, not shown in the figure, are provided for protecting the eye in observations of the sun, and a vernier and reading microscope are provided instead of the pointer I.

966. Spherical Mirrors.—By a spherical mirror is meant a mirror

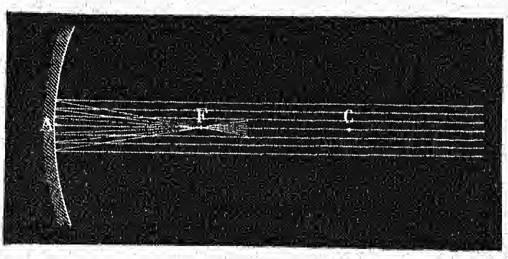


Fig. 673.—Principal Focus.

whose reflecting surface is a portion (usually a very small portion) of the surface of a sphere. It is concave or convex according as the inside or outside of the spherical surface yields the reflection. The centre of the sphere (C, Fig. 673) is called the *centre of curvature* of

the mirror. If the mirror has a circular boundary, as is usually the case, the central point A of the reflecting surface may conveniently be called the *pole* of the mirror. *Centre of the mirror* is an ambiguous phrase, being employed sometimes to denote the pole, and sometimes the centre of curvature. The line A C is called the *principal axis* of the mirror, and any other straight line through C which meets the mirror is called a *secondary axis*.

When the incident rays are parallel to the principal axis, the reflected rays converge to a point F, which is called the *principal focus*. This law is rigorously true for parabolic mirrors (generated by the revolution of a parabola about its principal axis). For spherical mirrors it is only approximately true, but the approximation is very close if the mirror is only a very small portion of an entire sphere. In grinding and polishing the specula of large reflecting telescopes, the attempt is made to give them, as nearly as possible, the parabolic form. Parabolic mirrors are also frequently employed

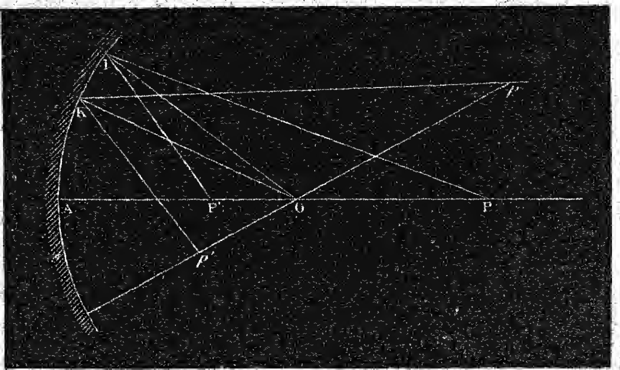


Fig. 674.—Theory of Conjugate Foci.

to reflect, in a definite direction, the rays of a lamp placed at the focus.

Rays reflected from the circumferential portion of a spherical mirror are always too convergent to concur exactly with those reflected from the central portion. This deviation from exact concurrence is called *spherical aberration*.

967. **Conjugate Foci.**—Let P (Fig. 674) be a luminous point situated on the principal axis of a spherical mirror, and let P I be one of the rays which it sends to the mirror. Draw the normal O I, which is simply a radius of the sphere. Then O I P is the angle of incid-

ence, and the angle of reflection OIP' must be equal to it; hence OI bisects an angle of the triangle PIP', and therefore we have

$$\frac{IP}{IP'} = \frac{OP}{OP'}$$

Let p, p' denote AP, AP' respectively, and let r denote the radius of the sphere. Then, if the angular aperture of the mirror is small, IP is sensibly equal to p, and IP' to p'. Substituting these approximate values, the preceding equation becomes

$$\frac{p}{p'} = \frac{p-r}{r-p'}; \text{ whence } pr + p'r = 2pp';$$

or, dividing by $pp'r$,

$$\frac{1}{p} + \frac{1}{p'} = \frac{2}{r}. \qquad (\alpha)$$

This formula determines the position of the point P', in which the reflected ray cuts the principal axis, and shows that it is, to the accuracy of our approximation, independent of the position of the point I; that is to say, all the rays which P sends to the mirror are reflected to the same point P'. We have assumed P to be on the principal axis. If we had taken it on a secondary axis, as at p (Fig. 674), we should have found, by the same process of reasoning, that the reflected rays would all meet in a point p' on that secondary axis. The distinction between primary and secondary axes, in the case of a spherical mirror, is in fact merely a matter of convenience, not representing any essential difference of property. Hence we can lay down the following general proposition as true within limits of error corresponding to the approximate equalities which we have above assumed as exact:—

Rays proceeding from any given point in front of a concave spherical mirror, are reflected so as to meet in another point; and the line joining the two points passes through the centre of the sphere.

It is evident that rays proceeding from the second point to the mirror, would be reflected to the first. The relation between them is therefore mutual, and they are hence called *conjugate foci.* By a *focus* in general is meant a point in which a number of rays, which originally came from the same point, meet (or would meet if produced); and the rays which thus meet, taken collectively, are called a *pencil.* Fig. 675 represents two pencils of rays whose foci S *s* are conjugate, so that, if either of them be regarded as an incident pencil, the other will be the corresponding reflected pencil.

We can now explain the formation of images by concave mirrors. Each point of the object sends a pencil of rays to the mirror, which converge, after reflection, to the conjugate focus. If the eye of the observer be placed beyond this point of concourse, and in the path of the rays, they will present to him the same appearance as if they

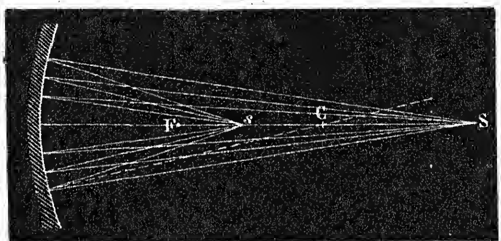


Fig. 675.—Conjugate Foci.

had come from this point as origin. The image is thus composed of points which are the conjugate foci of the several points of the object.

968. Principal Focus.—If, in formula (a) of last section, we make p increase continually, the term $\frac{1}{p}$ will continually decrease, and will vanish as p becomes infinite. This is the case of rays parallel to the principal axis, for parallel rays may be regarded as coming from a point at infinite distance. The formula then becomes

$$\frac{1}{p'} = \frac{2}{r}; \text{ whence } p' = \frac{r}{2};$$

that is to say, *the principal focal distance is half the radius of curvature.* This distance is often called the *focal length* of the mirror. If we denote it by f, the general formula becomes

$$\frac{1}{p} + \frac{1}{p'} = \frac{1}{f}. \qquad (b)$$

969. Discussion of the Formula.—By the aid of this formula we can easily trace the corresponding movements of conjugate foci.

If p is positive and very large, p' is a very little greater than f; that is to say, the conjugate focus is a very little beyond the principal focus.

As p diminishes, p' increases, until they become equal, in which case each of them is equal to r or $2f$; that is to say, the conjugate foci move towards each other till they coincide at the centre of curvature. This last result is obvious in itself; for rays from the centre

of curvature are normal to the mirror, and are therefore reflected directly back.

As p continues to diminish, the two foci, as it were, change places; the luminous point advancing from the centre of curvature to the principal focus, while the conjugate focus moves away from the centre of curvature to infinity.

As the luminous point continues to approach the mirror, $\frac{1}{p}$ is greater than $\frac{1}{f}$, and hence $\frac{1}{p'}$, and therefore also p', must be negative. The physical interpretation of this result is that the conjugate focus is *behind* the mirror, as at *s* (Fig. 676), and that the reflected

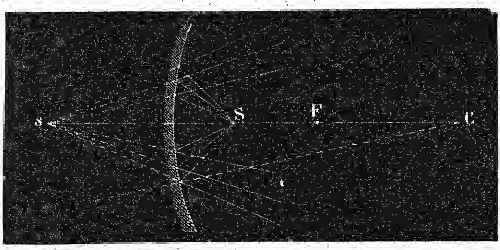


Fig. 676.—Virtual Focus.

rays diverge as if they had come from this point. Such a focus is called *virtual*, while a focus in which rays actually meet is called *real*. As the luminous point moves up from F to the mirror, the conjugate focus moves up from an infinite distance at the back, and meets it at the surface of the mirror.

If S is a real luminous point sending rays to the mirror, it must of necessity lie in front of the mirror, and p therefore cannot be negative; but when we are considering images of images this restriction no longer holds. If an incident beam, for example, converges towards a point *s* at the back of the mirror, it will be reflected to a point S in front. In this case p is negative, and p' positive. The conjugate foci S *s* have in fact changed places.

It appears from the above investigation that there are two principal cases, as regards the positions of conjugate foci of a concave mirror.

1. One focus between F and C; and the other beyond C.

2. One focus between F and the mirror; and the other behind the mirror.

In the former case, the foci move to meet each other at C; in the latter, they move to meet each other at the surface of the mirror.

957-967-968-

970. Formation of Images.—We are now in a position to discuss the formation of images by concave mirrors. Let A B (Fig. 677) be an object placed in front of a concave mirror, at a distance greater than its radius of curvature. All the rays which diverge from A will be reflected to the conjugate focus a. Hence this point can be found by the following construction. Draw through A the ray A A′ parallel to the principal axis, and draw its path after reflection, which must of necessity pass through the principal focus. The intersection of this reflected ray with the secondary axis through A will be the point required. A similar construction will give the conjugate focus

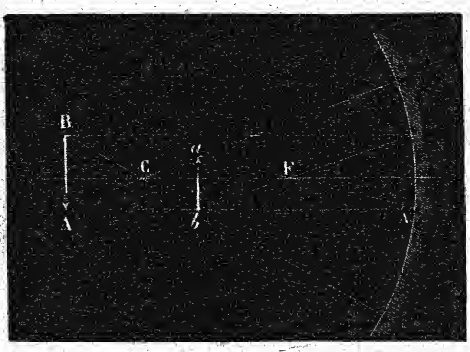


Fig. 677.—Formation of Image.

corresponding to any other point of the object; b, for example,[1] is the focus conjugate to B. Points of the object lying between A and B will have their conjugate foci between a and b. An eye placed behind the object A B will accordingly receive the same impression from the reflected rays as if the image $a\,b$ were a real object.

Since the lines joining corresponding points of object and image cross at the point C, which lies between them when the image is real, a real image formed by a concave mirror is always inverted.

971. Size of Image.—As regards the comparative sizes of object and image, it is obvious, from similar triangles, that their linear dimensions are *directly as their distances from* C, *the centre of curvature.*

Again, we have proved in § 967 that, in the notation of that section,

$$\frac{\text{I P}}{\text{I P}'} = \frac{\text{O P}}{\text{O P}'};$$

[1] It is only by accident that b happens to lie on A A′ in the figure.

or, the distances of object and image from the mirror are directly as their distances from the centre of curvature. Their linear dimensions are therefore *directly as their distances from the mirror.*

Again, by equation (*b*),

$$\frac{1}{p'} = \frac{1}{f} - \frac{1}{p} = \frac{p-f}{fp},$$

whence

$$\frac{p}{p'} = \frac{p-f}{f}, \qquad (c)$$

where $p-f$ is the distance of the object from the principal focus. Hence the linear dimensions of object and image are *in the ratio of the distance of the object from the principal focus to the focal length.*

These three rules are perfectly general, both for concave and convex mirrors.

The first rule shows that the object and image are equal when

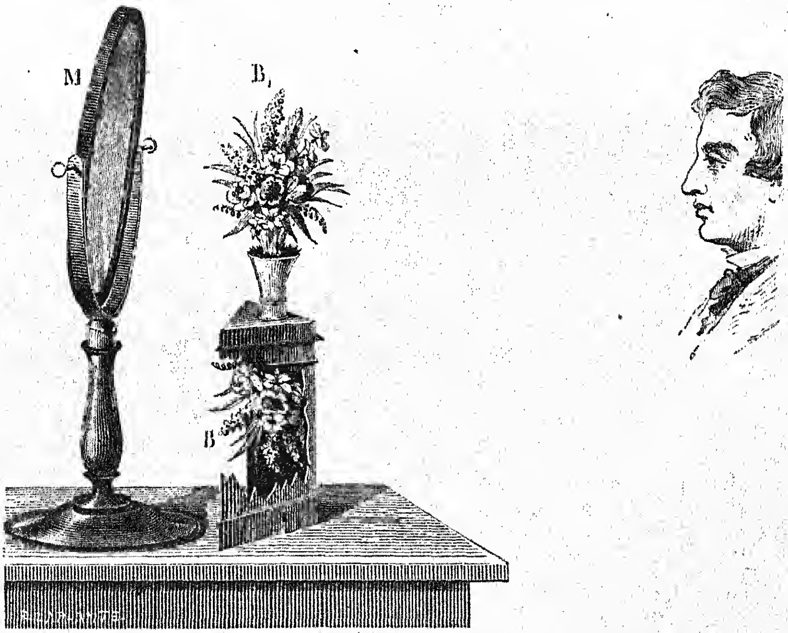


Fig. 678.—Experiment of Phantom Bouquet.

they coincide at the reflecting surface, and that, as they separate from this point in opposite directions, that which moves away from the centre of curvature continually gains in size upon the other.

The second rule shows that the object and image are equal when

they coincide at the centre of curvature, and that as they separate from this point, in opposite directions, that which moves away from the mirror continually gains in size upon the other.

The third rule shows that, when the object is at the principal focus, the size of the image is infinite.

972. Experiment of the Phantom Bouquet.—Let a box, open on one side, be placed in front of a concave mirror (Fig. 678), at a distance about equal to its radius of curvature, and let an inverted bouquet be suspended within it, the open side of the box being next the mirror. By giving a proper inclination to the mirror, an image of the bouquet will be obtained in mid-air, just above the top of the box. As the bouquet is inverted, its image is erect, and a real vase may be placed in such a position that the phantom bouquet shall appear to be standing in it. The spectator must be full in front of the mirror, and at a sufficient distance for all parts of the image to lie between his eyes and the mirror. When the colours of the bouquet are bright, the image is generally bright enough to render the illusion very complete.

973. Images on a Screen.—Such experiments as that just described

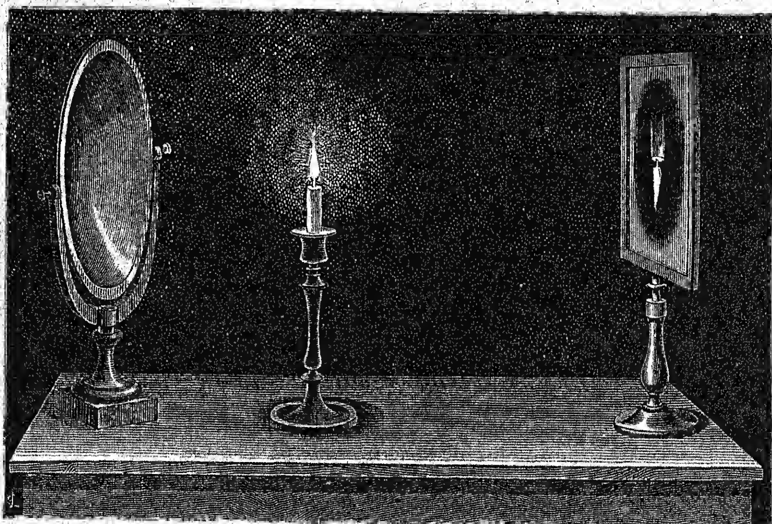


Fig. 679.—Image on Screen.

can only be seen by a few persons at once, since they require the spectator to be in a line with the image and the mirror. When an image is projected on a screen, it can be seen by a whole audience

at once, if the room be darkened and the image be large and bright. Let a lighted candle, for example, be placed in front of a concave mirror, at a distance exceeding the focal length, and let a screen be placed at the conjugate focus; an inverted image of the candle will be depicted on the screen. Fig. 679 represents the case in which the candle is at a distance less than the radius of curvature, and the image is accordingly magnified.

By this mode of operating, the formula for conjugate focal distances can be experimentally verified with considerable rigour, care being taken, in each experiment, to place the screen in the position which gives the most sharply defined image.

974. Difference between Image on Screen, and Image as seen in Mid-air. Caustics.—For the sake of simplicity we have made some statements regarding visible images which are not quite accurate; and we must now indicate the necessary corrections.

Images thrown on a screen have a determinate position, and are really the loci of the conjugate foci of the points of the object; but

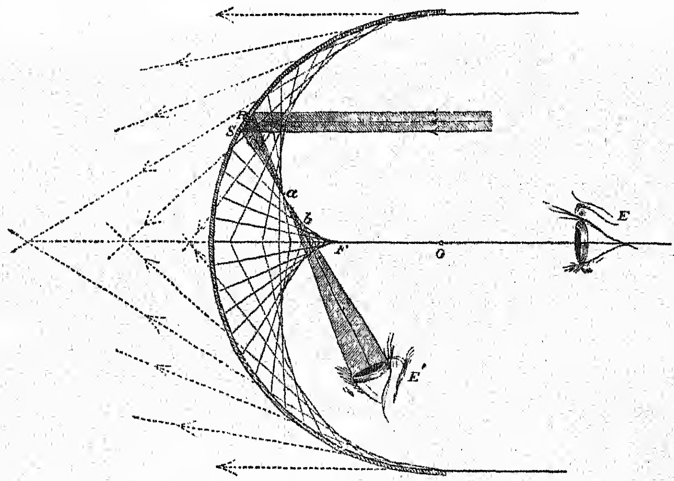


Fig. 680.—Position of Image in Oblique Reflection.

this is not rigorously true of images seen directly. They change their position to some extent, according to the position of the observer.

The actual state of things is explained by Fig. 680. The plane of the figure[1] is a principal plane (that is, a plane containing the principal axis) of a concave hemispherical mirror, and the incident rays

[1] Figs. 680 and 698 are borrowed, by permission, from Mr. Osmund Airy's *Geometrical Optics.*

are parallel to the principal axis. All the rays reflected in the plane of the figure touch a certain curve called a *caustic curve*, which has a cusp at F, the principal focus; and the direction in which the image is seen by an eye situated in the plane of the figure is determined by drawing from the eye a tangent to this caustic. If the eye be at E, on the principal axis, the point of contact will be F; but when the rays are received obliquely, as at E′, it will be at a point a not lying in the direction of F. For an eye thus situated, a is called the *primary focus*, and the point where the tangent at a cuts the principal axis is called the *secondary focus*. When the eye is moved in the plane of the diagram, the apparent position of the image (as determined by its remaining in coincidence with a cross of threads or other mark) is the primary focus; and when the eye is moved perpendicular to the plane of the diagram, the apparent position of the image is the secondary focus.[1] If we suppose the diagram to rotate about the principal axis, it will still remain true in all positions, and the surface generated by this revolution of the caustic curve is the *caustic surface*. Its form and position vary with the position of the point from which the incident rays proceed; and it has a cusp at the focus conjugate to this point.

There is always more or less blurring, in the case of images seen obliquely (except in plane mirrors), by reason of the fact that the point of contact with the caustic surface is not the same for rays entering different parts of the pupil of the eye.

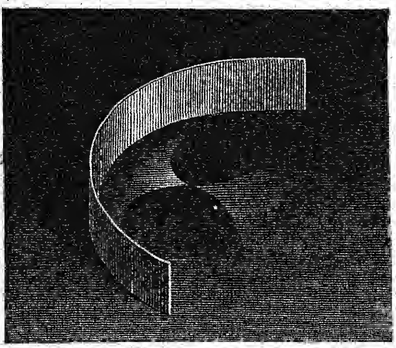


Fig. 681.—Caustic by Reflection.

A caustic curve can be exhibited experimentally by allowing the rays of the sun or of a lamp to fall on the concave surface of a strip of polished metal bent into the form of a circular arc, as in Fig. 681, the reflected light being received on a sheet of white paper on which the strip rests. The same effect may often be observed on the surface of a cup of tea, the reflector in this case being the inside of the tea-cup.

[1] Since every ray incident parallel to the principal axis, is reflected through the principal axis. If the incident rays diverged from a point on the principal axis, they would still be reflected through the principal axis.

The image of a luminous point received upon a screen is formed by all the rays which touch the corresponding caustic surface. The brightest and most distinct image will be formed at the cusp, which is, in fact, the conjugate focus; but there will be a border of fainter light surrounding it. This source of indistinctness in images is an example of *spherical aberration* (§ 967).

975. Image on a Screen by Oblique Reflection.—If we attempt to throw upon a screen the image of a luminous point by means of a concave mirror very oblique to the incident rays, we shall find that no image can be obtained at all resembling a point; but that there are two positions of the screen in which the image becomes a line.

In the annexed figure (Fig. 682), which represents on a larger scale a portion of Fig. 680, *a c*, *b d* are rays from the highest and lowest points of the portion R S of the hemispherical mirror, which portion we suppose to be small in both its dimensions in comparison with the radius of curvature; and we may suppose the rest of the hemisphere to be removed, so that R S will represent a small concave mirror receiving a pencil very obliquely.

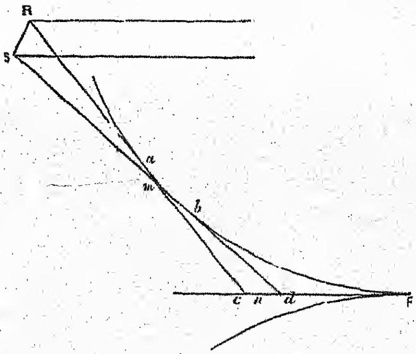


Fig. 682.—Formation of Focal Lines.

Then, if a screen be held perpendicular to the plane of the diagram, at *m*, where the section of the pencil by the plane of the diagram is narrowest, a blurred line of light will be formed upon it, the length of the line being perpendicular to the plane of the diagram. This is called the *primary focal line.*

The *secondary focal line* is *c d*, which, if produced, passes through the centre of curvature of the mirror, and also through the point from which the incident light proceeds. This line is very sharply formed upon a screen held so as to coincide with *c d* and to be perpendicular to the plane of the diagram. Its edges are much better defined than those of the primary line; and its position in space is also more definite. If the mirror is used as a burning-glass to collect

the sun's rays, ignition will be more easily obtained at one of these lines than in any intermediate position.

Focal lines can also be seen directly. In this case a small element of the mirror sends all its reflected rays to the eye, the rays from opposite sides of the element crossing each other at the focal lines, before they reach the eye. It is possible, in certain positions of the eye, to see either focal line at pleasure, by altering the focal adjustment of the eye; or the two may be seen with imperfect definition crossing each other at right angles. The experiment is easily made by employing a gas flame, turned very low, as the source of light. One line is in the plane of incidence, and the other is normal to this plane.

976. Virtual Image in Concave Mirror.—Let an object be placed, as in Figs. 683, 684, in front of a concave mirror, at a distance less than that of the principal focus. The rays incident on the mirror from any point of it, as A (Fig. 683), will be reflected as a divergent pencil, the focus from which they diverge being a point b at the back of the mirror. To find this point, we may trace the course of a ray through A parallel to the principal axis. Such a ray will be reflected to the principal focus F, and by producing this reflected ray backwards till it meets the secondary axis C A, the point b, which is the conjugate focus of A, is determined. We can find in the same way the position of a, the conjugate focus of B, and it is obvious that the image of A B will be erect and magnified.

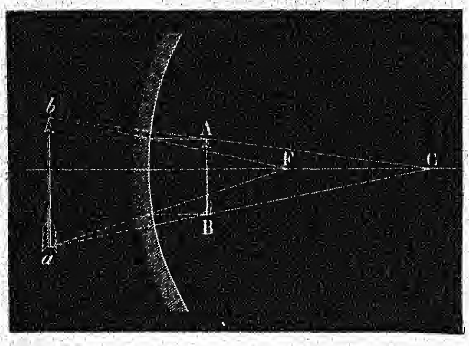


Fig. 683.—Formation of Virtual Image.

977. Remarks on Virtual Images.—A virtual image cannot be projected on a screen; for the rays which produce it do not actually pass through its place, but only seem to do so. A screen placed at $a\,b$ would obviously receive none of the reflected light whatever.

[1] The "elongated figure of 8" which is often mentioned in connection with the secondary focal line, is obtained by turning the screen about n the middle point of $c\,d$, so as to blur both ends of the image by bad focussing. It will be observed, from an inspection of the diagram, that $c\,d$ is very oblique to the reflected rays.

If we neglect the blurring of the primary line, we may describe the part of the pencil lying between the two lines as a tetrahedron, of which the two lines are opposite edges.

The images seen in a plane mirror are virtual; and any spherical mirror, whether concave or convex, is nearly equivalent to a plane mirror, when the distance of the object from its surface is small in comparison with the radius of curvature.

978. Convex Mirrors.—It is easily shown, by a simple construction, that rays incident from any luminous point upon a convex mirror, diverge after reflection. The principal focus, and the foci conjugate to all points external to the sphere, are therefore virtual.

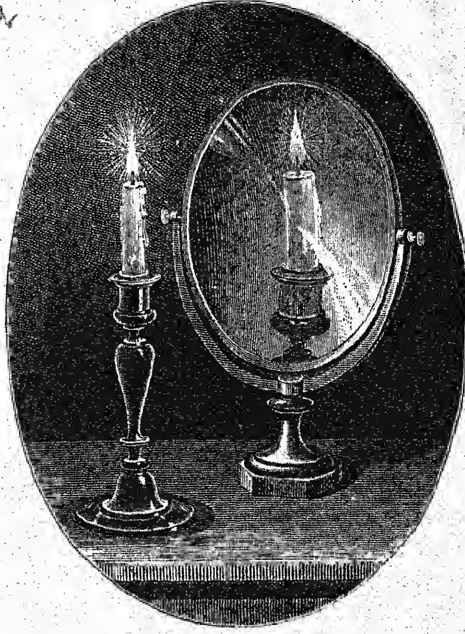


Fig. 684.—Virtual Image in Concave Mirror.

To adapt formulæ (a) and (b) of the preceding sections to the case of convex mirrors, we have only to alter the sign of the term $\frac{2}{r}$ or $\frac{1}{f}$; so that for a convex mirror we shall have

$$\frac{1}{p}+\frac{1}{p'}=-\frac{1}{f}=-\frac{2}{r}; \qquad (c)$$

r and f being here regarded as essentially positive.

From this formula it is obvious that one at least of the two distances p, p' must be negative; that is to say, one at least of any pair of conjugate foci must lie behind the mirror.

The construction for an image (Fig. 685) is the same as in the case of concave mirrors. Through any selected point of the object draw a ray parallel to the principal axis; the reflected ray, if produced backwards, must pass through the principal focus, and its intersection with the secondary axis through the selected point determines the corresponding point of the image. The image of an external object will evidently be erect, and smaller than the object. Repeating the same construction when the object is nearer to the mirror, we see that the image will be larger than before.

The linear dimensions of an object and its image, whether in the case of a convex or a concave mirror, are directly proportional to

their distances from the centre of curvature, and are also directly proportional to their distances from the mirror. The image is inverted or erect according as the centre of curvature does or does not

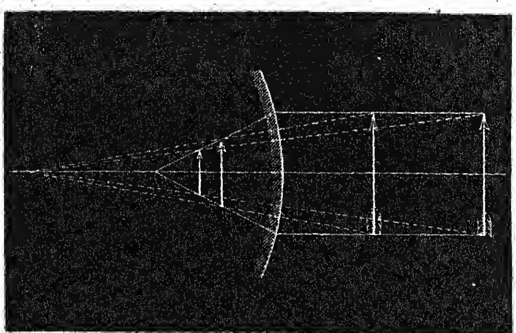


Fig. 685.—Formation of Image in Convex Mirror.

lie between the object and its image. In the case of a convex mirror the centre never lies between them (if the object be real), and therefore the image is always erect.

Convex mirrors are very seldom employed in optical instruments.

The silvered globes which are frequently used as ornaments, are examples of convex mirrors, and present to the observer at one view an image of nearly the whole surrounding landscape. As the part of the mirror in which he sees this image is nearly an entire hemisphere, the deformation of the image is very notable, straight lines being reflected as curves.

979. **Anamorphosis.**—Much greater deformations are produced by cylindric mirrors. A cylindric mirror, when the axis of the cylinder is vertical, behaves like a plane mirror as regards the angular magnitude under which the height of the image is seen, and like a spherical mirror as regards the breadth of the image. If it be a convex cylinder, it causes bodies to appear unduly contracted horizontally in proportion to their heights. Distorted pictures are sometimes drawn upon paper, according to such a system that when they are seen reflected in a cylindric mirror properly placed, as in Fig. 686, the distortion is corrected, and while the picture appears a mass of confusion, the image is instantly recognized. This restoration of true proportion in a picture is called *anamorphosis.*

980. **Medical Applications.**—Concave mirrors are frequently used

to concentrate light upon an object for the purpose of rendering it more distinctly visible.

The *ophthalmoscope* is a small concave mirror, with a small hole in its centre, through which the observer looks from behind, while he

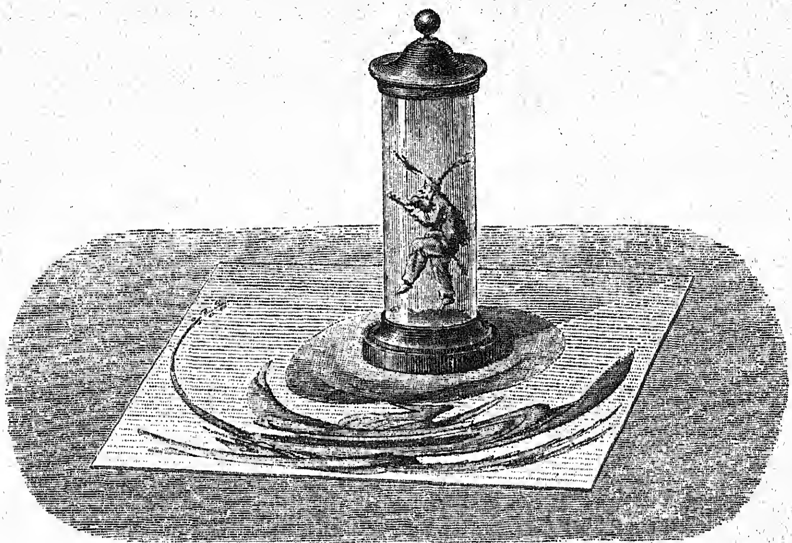


Fig. 686.—Anamorphosis.

directs a beam of reflected light from a lamp into the pupil of the patient's eye. In this way (with the help sometimes of a lens) the retina can be rendered visible, and can be minutely examined.

The *laryngoscope* consists of two mirrors. One is a small plane mirror, with a handle attached, at an angle of about 45° to its plane. This small mirror is held at the back of the patient's mouth, so that the observer, looking into it, is able by reflection to see down the patient's throat, the necessary illumination being supplied by a concave mirror, strapped to the observer's forehead, by means of which the light from a lamp is reflected upon the plane mirror, which again reflects it down the throat.

CHAPTER LXIX.

REFRACTION.

981. **Refraction.**—When a ray of light passes from one transparent medium to another, it undergoes a change of direction at the surface of separation, so that its course in the second medium makes an angle with its course in the first. This changing of direction is called *refraction.*

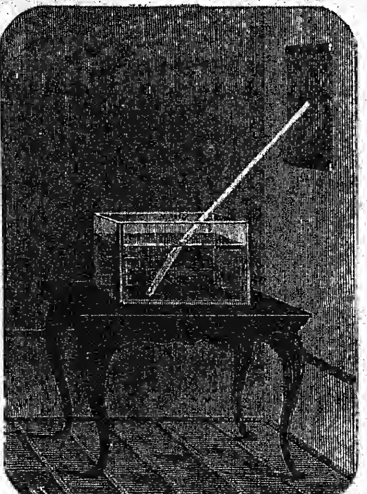


Fig. 687.—Refraction.

The phenomenon can be exhibited by admitting a beam of the sun's rays into a dark room, and receiving it on the surface of water contained in a rectangular glass vessel (Fig. 687). The path of the beam will be easily traced by its illumination of the small solid particles which lie in its course.

The following experiment is a well-known illustration of refraction:—A coin $m\,n$ (Fig. 688) is laid at the bottom of a vessel with opaque sides, and a spectator places himself so that the coin is just hidden from him by the side of the vessel; that is to say, so that the line m A in the figure passes just above his eye. Let water now be poured into the vessel, care being taken not to displace the coin. The bottom of the vessel will appear to rise, and the coin will come into sight. Hence a pencil of rays from m must have entered the spectator's eye. The pencil in fact undergoes a sudden bend at the surface of the water, and thus reaches the eye by a crooked course,

in which the obstacle A is evaded. If the part of the pencil in air be produced backwards, its rays will approximately meet in a point m', which is therefore the image of m. Its position is not correctly indicated in the figure, being placed too much to the left (§ 990).

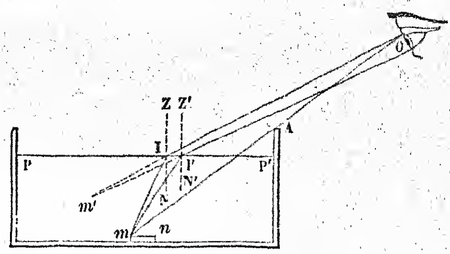


Fig. 688.—Experiment of Coin in Basin.

The broken appearance presented by a stick (Fig. 689) when partly immersed in water in an oblique position, is similarly explained, the part beneath the water being lifted up by refraction. Quit

982. Refractive Powers of Different Media.—In the experiments of the coin and stick, the rays, in leaving the water, are bent away from the normals Z I N, Z′ I′ N′ at the points of emergence; in the experiment first described (Fig. 687), on the other hand, the rays, in passing from air into water, are bent nearer to the normal. In every case the path which the rays pursue in going is the same as they would pursue in returning; and of the two media concerned, that in which the ray makes the smaller angle with the normal is said to have greater refractive power than the other, or to be more highly refracting.

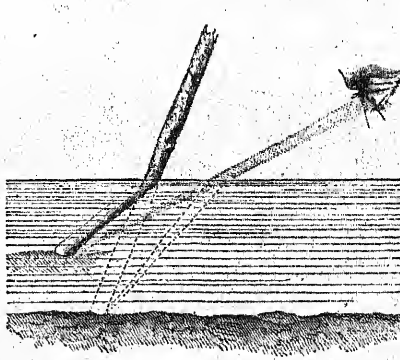


Fig. 689.—Appearance of Stick in Water.

Liquids have greater refractive power than gases, and as a general rule (subject to some exceptions in the comparison of dissimilar substances) the denser of two substances has the greater refracting power. Hence it has become customary, in enunciating some of the laws of optics, to speak of the *denser* medium and the *rarer* medium, when the more correct designations would be *more refractive* and *less refractive*.

983. Laws of Refraction.—The quantitative law of refraction was not discovered till quite modern times. It was first stated by Snell, a Dutch philosopher, and was made more generally known by Descartes, who has often been called its discoverer.

Let R I (Fig. 690) be a ray incident at I on the surface of separation of two media, and let I S be the course of the ray after refraction. Then the angles which R I and I S make with the normal are called the *angle of incidence* and the *angle of refraction* respectively; and the first law of refraction is that these angles lie in the same plane, or *the plane of refraction is the same as the plane of incidence.*

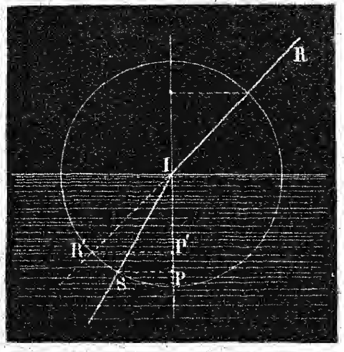


Fig. 690.—Law of Refraction.

The law which connects the magnitudes of these two angles, and which was discovered by Snell, can only be stated either by reference to a geometrical construction, or by employing the language of trigonometry. Describe a circle about the point of incidence I as centre, and drop perpendiculars, from the points where it cuts the rays, on the normal. The law is that these perpendiculars R′ P′, S P, will have a constant ratio; or *the sines of the angles of incidence and refraction are in a constant ratio.* It is often referred to as the *law of sines.*

The angle by which a ray is turned out of its original course in undergoing refraction is called its *deviation.* It is zero if the incident ray is normal, and always increases with the angle of incidence.

read
15 986

984. Verification of the Law of Sines.—These laws can be verified by means of the apparatus represented in Fig. 691, which is very similar to that employed by Descartes. It has a vertical divided circle, to the front of which is attached a cylindrical vessel, half-filled with water or some other transparent liquid. The surface of the liquid must pass exactly through the centre of the circle. I is a movable mirror for directing a reflected beam of solar light on the centre O. The beam must be directed centrally through a short tube attached to the mirror, and to facilitate this adjustment the tube is furnished with a diaphragm with a hole in its centre. The arm O *a* is movable about the centre of the circle, and carries a vernier for measuring the angle of incidence. The ray undergoes refraction at O; and the angle of refraction is measured by means of a second arm O R, which is to be moved into such a position that the diaphragm of its tube receives the beam centrally. No refraction

occurs at emergence, since the emergent beam is normal to the surfaces of the liquid and glass; the position of the arm accordingly indicates the direction of the refracted ray. The angles of incidence and refraction can be read off at the verniers carried by the two arms; and the ratio of their sines will be found constant. The sines can also be directly measured by employing sliding-scales as indicated in the figure, the readings being taken at the extremity of each arm.

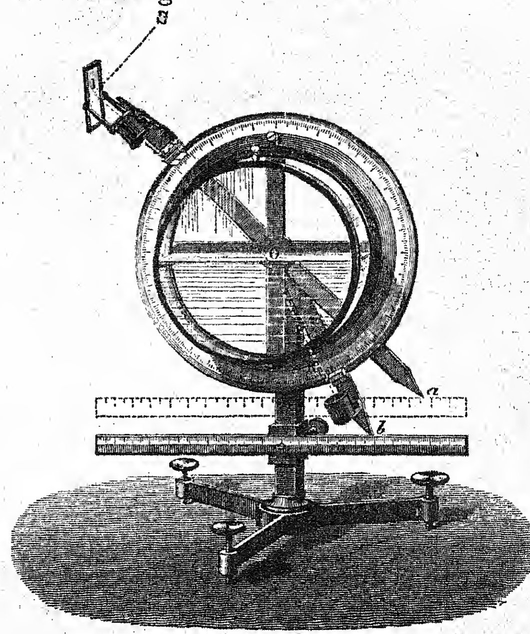


Fig. 691.—Apparatus for Verifying the Law.

It would be easy to make a beam of light enter at the lower side of the apparatus, in a radial direction; and it would be found that the ratio of the sines was precisely the same as when the light entered from above. This is merely an instance of the general law, that the course of a returning ray is the same as that of a direct ray.

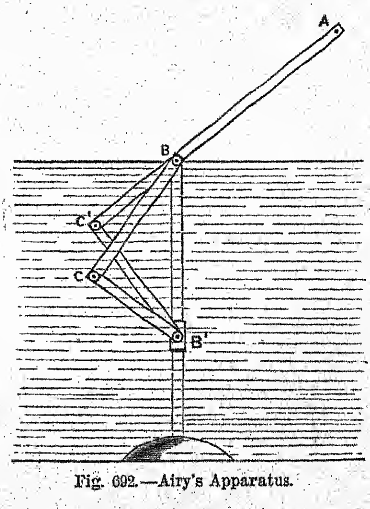


Fig. 692.—Airy's Apparatus.

985. Airy's Apparatus.—The following apparatus for the same purpose was invented, many years ago, by the present astronomer royal. B′ is a slider travelling up and down a vertical stem. A C′ and B C are two rods pivoted on a fixed point B of the vertical stem. C′ B′ and C B′ are two other rods jointed to the former at C′ and C, and pivoted at their lower ends on the centre

of the slider. BC is equal to $B'C'$, and BC' to $B'C$. Hence the two triangles BCB', $B'C'B$ are equal to one another in all positions of the slider, their common side BB' being variable, while the other two sides of each remain unchanged in length though altered in position.

The ratio $\frac{BC}{CB'}$ or $\frac{B'C'}{C'B}$ is made equal to the index of refraction of the liquid in which the observation is to be made. For water this ratio will be $\frac{4}{3}$. Then, if the apparatus is surrounded with water up to the level of B, ABC will be the path of a ray, and a stud at C will appear in the same line with studs at A and B; for we have

$$\frac{\sin C'BB'}{\sin CBB'} = \frac{\sin C'BB'}{\sin C'B'B} = \frac{C'B'}{C'B} = \frac{4}{3}.$$

986. Indices of Refraction.—The ratio of the sine of the angle of incidence to the sine of the angle of refraction when a ray passes from one medium into another, is called the *relative index of refraction* from the former medium to the latter. When a ray passes from vacuum into any medium this ratio is always greater than unity, and is called the *absolute index of refraction*, or simply the *index of refraction*, for the medium in question. The relative index of refraction from any medium A into another B is always equal to the absolute index of B divided by the absolute index of A. The absolute index of air is so small that it may usually be neglected in comparison with those of solids and liquids: but strictly speaking, the relative index for a ray passing from air into a given substance must be multiplied by the absolute index for air, in order to obtain the absolute index of refraction for the substance.

The following table gives the indices of refraction of several substances:—

INDICES OF REFRACTION.[1]

Diamond,	2·44 to 2·755	Alcohol,	1·372
Sapphire,	1·794	Aqueous humour of eye,	1·337
Flint-glass,	1·576 to 1·642	Vitreous humour,	1·339
Crown-glass,	1·531 to 1·563	Crystalline lens, outer coat,	1·337
Rock-salt,	1·545	„ „ under coat,	1·379
Canada balsam,	1·540	„ „ central portion,	1·400
Bisulphide of carbon,	1·678	Sea water,	1·343
Linseed oil (sp. gr. ·932),	1·482	Pure water,	1·336
Oil of turpentine (sp. gr. ·885),	1·478	Air at 0° C. and 760 mm	1·000294

987. Critical Angle.—We see, from the law of sines, that when the

[1] The index of refraction is always greater for violet than for red (see Chap. lxxii.). The numbers in this table are to be understood as mean values.

incident ray is in the less refractive of the two media, to every possible angle of incidence there is a corresponding angle of refraction. This, however, is not the case when the incident ray is in the more refractive of the two media. Let S O, S′ O, S″ O (Fig. 693) be inci-

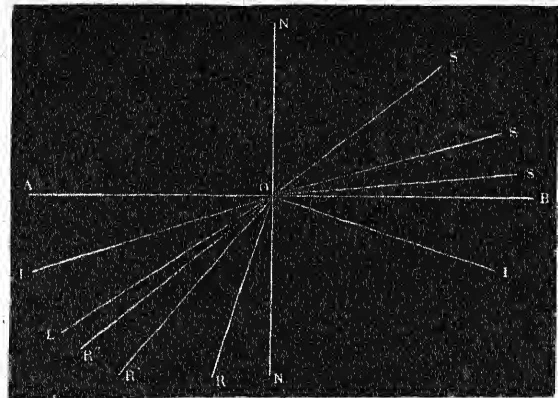


Fig. 693.—Critical Angle.

dent rays in the less refractive medium, and O R, O R′, O R″ the corresponding refracted rays. There will be a particular direction of refraction O L corresponding to the angle of incidence of 90°. Conversely, incident rays R O, R′ O, R″ O, in the more refractive medium, will emerge in the directions O S, O S′, O S″, and the direction of emergence for the incident ray L O will be O B, which is coincident with the bounding surface.

The angle L O N is called the critical angle, and is easily computed when the relative index of refraction is given. For let μ denote this index (the incident ray being supposed to be in the less refractive medium), then we are to have

$$\frac{\sin 90^\circ}{\sin x} = \mu, \text{ whence } \sin x = \frac{1}{\mu};$$

that is, *the sine of the critical angle is the reciprocal of the index of refraction.*

When the media are air and water, this angle is about 48° 30′. For air and different kinds of glass its value ranges from 38° to 41°.

If a ray, as I O, is incident in the more refractive medium, at an angle greater than the critical angle, the law of sines becomes nugatory, and experiment shows that such a ray undergoes internal reflection in the direction O I′, the angle of reflection being equal to

the angle of incidence. Reflection occurring in these circumstances is nearly perfect, and has received the name of *total reflection. Total reflection occurs when rays are incident in the more refractive medium at an angle greater than the critical angle.*

The phenomenon of total reflection may be observed in several familiar instances. For example, if a glass of water, with a spoon in it (Fig. 694), is held above the level of the eye, the under side of

Fig. 694.—Total Reflection.

the surface of the water is seen to shine like a brilliant mirror, and the lower part of the spoon is seen reflected in it. Beautiful effects of the same kind may be observed in aquariums.

938. Camera Lucida.—The *camera lucida* is an instrument sometimes employed to facilitate the sketching of objects from nature. It acts by total reflection, and may have various forms, of which that proposed by Wollaston, and represented in Figs. 695, 696, is

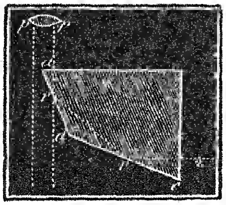


Fig. 695.—Section of Prism.

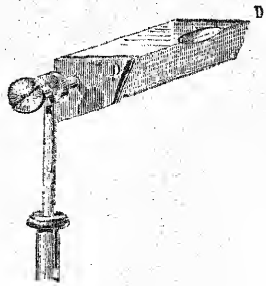


Fig. 696.—Camera Lucida.

one of the commonest. The essential part is a totally-reflecting prism with four angles, one of which is 90°, the opposite one 135°, and the other two each 67° 30′. One of the two faces which contain the right angle is turned towards the objects to be sketched. Rays incident normally on this face, as $x r$, make an angle greatly exceeding the critical angle with the face $c d$, and are totally reflected from it to the next face $d a$, whence they are again totally reflected to the fourth face, from which they emerge normally.[1] An eye placed so as to receive the emergent rays will see a virtual image in a direction at right angles to that in which the object lies. In practice, the eye is held over the angle a of the prism, in such a position that one-half of the pupil receives these reflected rays, while the other half receives light in a parallel direction outside the prism. The observer thus sees the reflected image projected on a real back-ground, which consists of a sheet of paper for sketching. He is thus enabled to pass a pencil over the outlines of the image; pencil, image, and paper being simultaneously visible. It is very desirable that the image should lie in the plane of the paper, not only because the pencil point and the image will then be seen with the same focussing of the eye, but also because parallax is thus obviated, so that when the observer shifts his eye the pencil point is not displaced on the image. A concave lens, with a focal length of something less than a foot, is therefore

[1] The use of having *two* reflections is to obtain an erect image. An image obtained by one reflection would be upside down.

placed close in front of the prism, in drawing distant objects. By raising or lowering the prism in its stand (Fig. 696), the image of the object to be sketched may be made to coincide with the plane of the paper.

The prism is mounted in such a way that it can be rotated either about a horizontal or a vertical axis; and its top is usually covered with a movable plate of blackened metal, having a semi-circular notch at one edge, for the observer to look through.

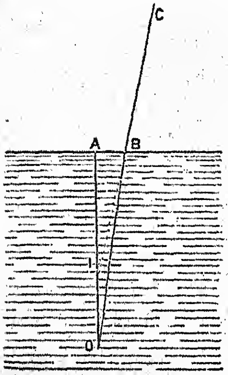


Fig. 697.—Image by Refraction.

989. Images by Refraction at a Plane Surface.—Let O (Fig. 697) be a small object in the interior of a solid or liquid bounded by a plane surface A B. Let O B C be the path of a nearly normal ray, and let B C (the portion in air) be produced backwards to meet the normal in I. Then, since A I B and A O B are the inclinations of the two portions of the ray to the normal, we have (if μ be the index of refraction from air into the substance)—

$$\mu = \frac{\sin AIB}{\sin AOB} = \frac{OB}{IB}.$$

But O B is ultimately equal to O A, and I B to I A. Hence, if we make A I equal to $\frac{AO}{\mu}$, all the emergent rays of a small and nearly normal pencil emitted by O will, if produced backwards, intersect O A at points indefinitely near to the point I thus determined. If the eye of an observer be situated on the production of the normal O A, the rays by which he sees the object O constitute such a pencil. He accordingly sees the image at I. As the value of μ is $\frac{4}{3}$ for water, and about $\frac{3}{2}$ for glass, it follows that the apparent depth of a pool of clear water when viewed vertically is $\frac{3}{4}$ of the true depth, and that the apparent thickness of a piece of plate-glass when viewed normally is only $\frac{2}{3}$ of the true thickness.

990.—When the incident pencil (Fig. 698) is not small, but includes rays of all obliquities, those of them which make angles with the normal less than the critical angle N Q R will emerge into air; and the emergent rays, if produced backwards, will all touch a certain

caustic surface, which has the normal Q N for its axis of revolution, and touches the surface at all points of a circle of which N R is the

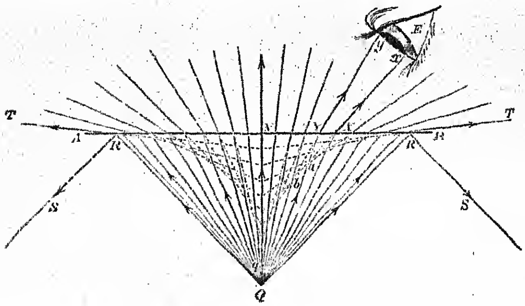


Fig. 698.—Caustic by Refraction.

radius. Wherever the eye may be situated, a tangent drawn from it to the caustic will be the direction of the visible image. If the observer sees the image with both eyes, both being equidistant from the surface and also equidistant from the normal, the two lines of sight thus determined (one for each eye) will meet at a point on the normal, which will accordingly be the apparent position of the image. If, on the other hand, both eyes are in the same plane containing the normal, the two lines of sight will intersect at a point between the normal and the observer.

The image, whether seen with one eye or two, approaches nearer to the surface as the direction of vision becomes more oblique, and ultimately coincides with it. The apparent depth of water, which is only $\frac{3}{4}$ of the real depth when seen vertically, is accordingly less than $\frac{3}{4}$ when seen obliquely, and becomes a vanishing quantity as the direction of vision approaches to parallelism with the surface. The focus I determined in the preceding section is at the cusp of the caustic.

991. Parallel Plate.—Rays falling normally on a uniform transparent plate with parallel faces, keep their course unchanged; but this is not the case with rays incident obliquely. A ray S I (Fig. 699), incident at the angle S I N, is refracted in the direction I R. The angle of incidence at R is equal to the angle of refraction at I, and hence the angle of emergence S′ R N′ is equal to the original angle of incidence S I N. The emergent ray R S′ is therefore parallel to the incident ray S I, but is not in the same straight line with it.

Objects seen obliquely through a plate are therefore displaced from their true positions. Let S (Fig. 700) be a luminous point

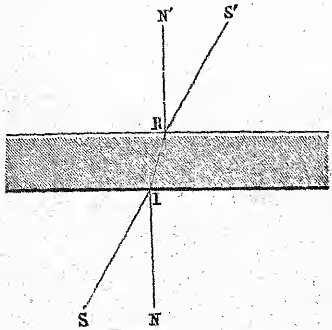


Fig. 699.—Parallel Plate.

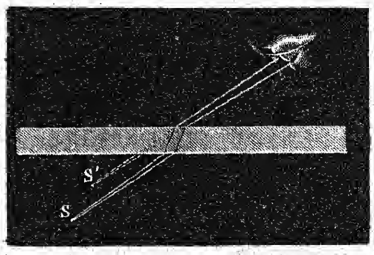


Fig. 700.—Vision through Plate.

which sends light to an eye not directly opposite to it, on the other side of a parallel plate. The emergent rays which enter the eye are parallel to the incident rays; but as they have undergone lateral displacement, their point of concourse[1] is changed from S to S', which is accordingly the image of S.

The displacement thus produced increases with the thickness of the plate, its index of refraction, and the obliquity of incidence. It furnishes one of the simplest means of measuring the index of refraction of a substance, and is thus employed in Pichot's refractometer.

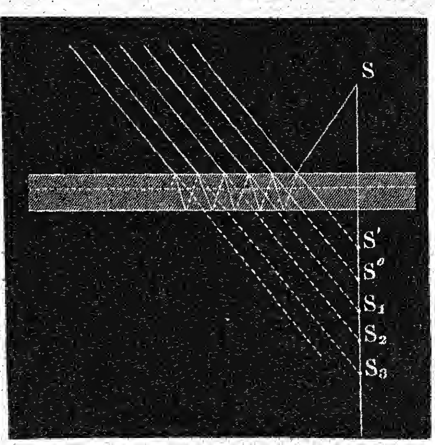


Fig. 701.—Multiple Images in Plate.

992. Multiple Images produced by a Plate.—Let S (Fig. 701) be a luminous point in front of a transparent plate with parallel faces. Of the rays which it sends to the plate, some will be reflected from the front, thus giving rise to an image S'. Another portion will enter the plate,

[1] The rays which compose the pencil that enters the eye will not exactly meet (when produced backwards) in any one point. There will be two focal lines, just as in the case of spherical mirrors (§ 974, 975).

undergo reflection at the back, and emerge with refraction at the front, giving rise to a second image S^0. Another portion will undergo internal reflection at the front, then again at the back, and by emerging in front will form a third image S_1. The same process may be repeated several times; and if the luminous object be a candle, or a piece of bright metal, a number of images, one behind another, will be visible to an eye properly placed in front (Fig. 702). All the successive images, after the first two, continually diminish in brightness. If the glass be silvered at the back, the second image is much brighter than the first, when the incidence is nearly normal, but as the angle of incidence increases, the first image gains upon the second, and ultimately surpasses it. This is due to the fact that the reflecting power of a surface of glass increases with the angle of incidence.

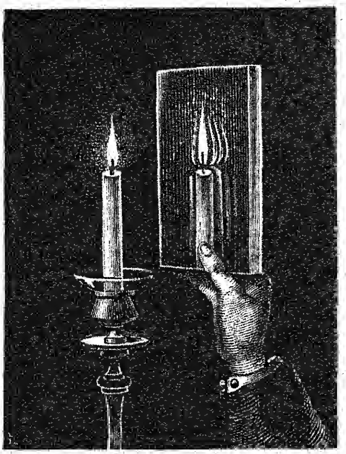


Fig. 702.
Images of Candle in Looking-glass.

If the luminous body is at a distance which may be regarded as infinite,—if it is a star, for example,—all the images should coincide, and form only a single image, occupying a position which does not vary with the position of the observer, provided that the plate is perfectly homogeneous, and its faces perfectly plane and parallel. A severe test is thus furnished of the fulfilment of these conditions.

Plates are sometimes tested, for parallelism and uniformity, by supporting them in a horizontal position on three points, viewing the image of a star in them with a telescope furnished with cross wires, and observing whether the image is displaced on the wires when the plate is shifted into a different position, still resting on the same three points.

993. Superimposed Plates. Astronomical Refraction.—We have stated in § 986 that the relative index from one medium into another is equal to the absolute index of the second divided by that of the first. Hence if μ_1 μ_2 are the absolute indices, and ϕ_1 ϕ_2 the angles which the two parts of the refracted ray make with the normal, we have

$$\sin \phi_1 = \frac{\mu_2}{\mu_1} \sin \phi_2,$$

or

$$\mu_1 \sin \phi_1 = \mu_2 \sin \phi_2. \qquad (1)$$

When a number of plates are superimposed, they will have a common normal. Let a ray pass through them all; let μ denote the absolute index of any one of the plates, and ϕ the angle which the portion of the ray that lies in this plate makes with the normal; then equation (1) shows that $\mu \sin \phi$ will have the same value for all parts of the ray. Hence if the value of ϕ for the first plate be given, its value for any plate in the series depends only on the value of μ for that plate, and will not be altered by removing some or all of the intervening plates.

This reasoning can be applied to the transmission of a ray from a star through the earth's atmosphere, if the distance of the star from the zenith does not exceed 20° or 30°. The portion of atmosphere traversed may be regarded as a series of horizontal plates, and the slope of the ray in the lowest plate will be the same as if all the plates above it were removed. In the case of stars near the horizon, the length of the path in air is so great that the curvature of the earth cannot be left out of account, in other words, the layers traversed cannot be regarded as parallel plates.

994. Refraction through a Prism.—For optical purposes, any portion of a transparent body lying between two plane faces which are not parallel may be regarded as a prism.[1] The line in which these faces meet, or would meet if produced, is called the edge of the prism, and a section made by a plane perpendicular to them both is called a *principal section.* The prisms chiefly employed are really prisms in the geometrical sense of the word. Their principal sections are usually triangular, and are very frequently equilateral, as in Fig. 703. The stand usually employed for prisms when mounted separately is represented in Fig. 704. It contains several joints. The uppermost is for rotating the prism about its own axis. The second is for turning the prism so that its edges shall make any required angle with the vertical. The third gives motion about a vertical axis, and also furnishes the means of raising and lowering the prism through a range of several inches.

Let S I (Fig. 705) be an incident ray in the plane of a principal section of the prism. If the external medium be air, or any other

[1] This amounts to saying that the word *prism* in optics means *wedge.*

substance of less refractive power than the prism, the ray in entering the prism will be bent nearer to the normal, taking such a course as I E, and in leaving the prism will be bent away from the normal,

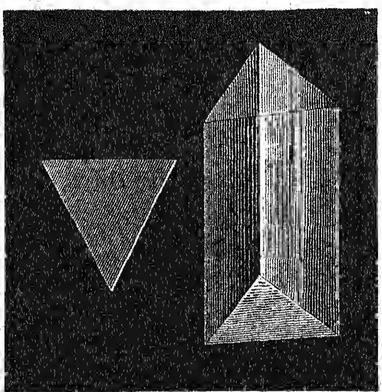


Fig. 703.—Equilateral Prism.

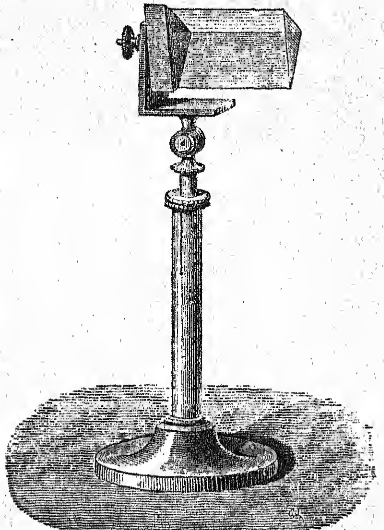


Fig. 704.—Prism mounted on Stand.

taking the course E B. The effect of these two refractions is, therefore, to turn the ray away from the edge (or refracting angle) of the prism. In practice, the prism is usually so placed that I E, the path of the ray through the prism, makes equal angles with the two faces at which refraction occurs (§ 995). If the prism is turned very far from this position, the course of the ray may be altogether different from that represented in the figure; it may, for example, enter at one face, be internally reflected at another, and come out at the third; but we at present exclude such cases from consideration.

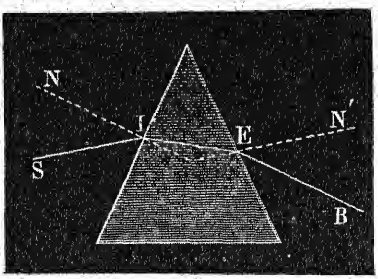


Fig. 705.—Refraction through Prism.

The direction of deviation is easily shown experimentally, by admitting a narrow beam of sunlight into a dark room, and introducing a prism in its course. It will be found that the refracted

beam, in the circumstances represented in Fig. 705, is turned aside some 40° or 50° from its original course.[1]

Since the rays which traverse a prism are bent away from the edge, the object from which they proceed will appear, to an observer looking through the prism, to be more nearly in the direction of the

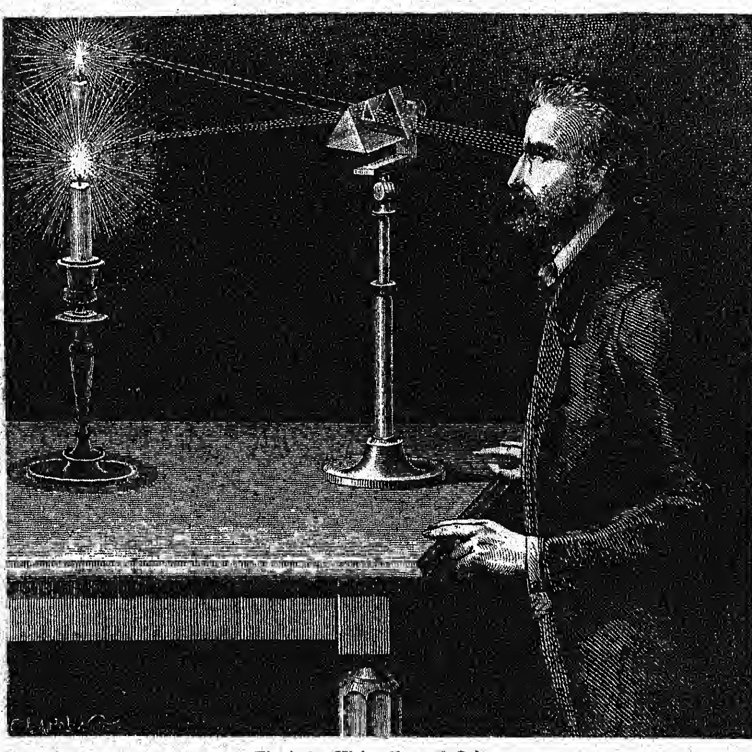


Fig. 706.—Vision through Prism.

edge than it really is. If, for example, he looks at the flame of a candle through a prism placed so that the edge which corresponds to the refracting angle is at the top (Fig. 706), the apparent place of the flame will be above its true place.

995. Formulæ for Refraction through Prisms. Minimum Deviation.—Let S I (Fig. 707) be an incident ray in the plane of a principal section A B C of a prism. Let i be the angle of incidence S I N, and

[1] The phenomena here described are complicated in practice by the unequal refrangibility of rays of different colours (Chap. lxxii.). The complication may be avoided by employing homogeneous light, of which a spirit-lamp, with common salt sprinkled on the wick, affords a nearly perfect example.

r the angle of refraction M I I′. Then, denoting the index of refraction by μ, we have $\sin i = \mu \sin r$. In like manner, putting r' for the angle of internal incidence on the second face I I′ M, and i' for the angle of external refraction N′ I′ R, we have $\sin i' = \mu \sin r'$.

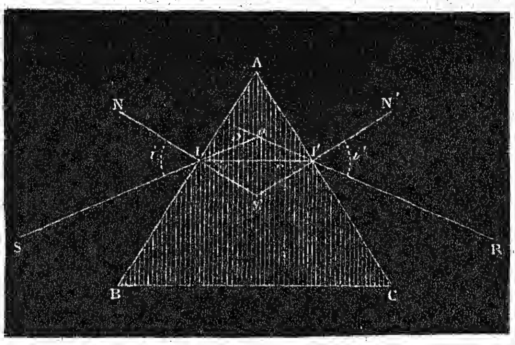


Fig. 707.—Refraction through Prism.

The deviation produced at I is $i - r$, and that at I′ is $i' - r'$, so that the total deviation, which is the acute angle D contained between the rays S I, R I′, when produced to meet at o, is

$$D = i - r + i' - r'. \tag{1}$$

But if we drop a perpendicular from the angular point A on the ray I I′, it will divide the refracting angle B A C into two parts, of which that on the left will be equal to r, and that on the right to r', since the angle contained between two lines is equal to that contained between their perpendiculars. We have therefore $A = r + r'$, and by substitution in the above equation

$$D = i + i' - A. \tag{2}$$

When the path of the ray through the prism I I′ makes equal angles with the two faces, the whole course of the ray is symmetrical with respect to a plane bisecting the refracting angle, so that we have

$$i = i'; \ r = r' = \frac{A}{2}.$$

Equation (2) thus becomes

$$D = 2\,i - A, \text{ whence } i = \frac{A + D}{2}, \tag{3}$$

$$\text{and } \mu = \frac{\sin i}{\sin r} = \frac{\sin \dfrac{A + D}{2}}{\sin \dfrac{A}{2}}. \tag{4}$$

This last result is of great practical importance, as it enables us to

calculate the index of refraction μ from measurements of the refracting angle A of the prism, and of the deviation D which occurs when the ray passes symmetrically.

When a beam of sunlight in a dark room is transmitted through a prism, it will be found, on rotating the prism about its axis, that there is a certain mean position which gives smaller deviation of the transmitted light than positions on either side of it; and that, when the prism is in this position, a small rotation of it has no sensible effect on the amount of deviation. The position determined experimentally by these conditions, and known as the *position of minimum deviation*, is the position in which the ray passes symmetrically.

996. Construction for Deviation.—The following geometrical construction furnishes a very simple method of representing the variation of deviation with the angle of incidence:—

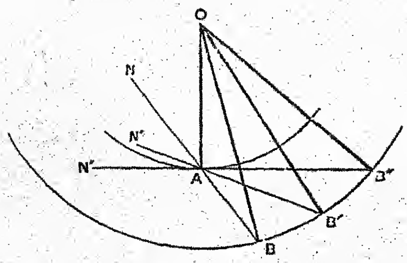


Fig. 708.
General Construction for Deviation.

1. When the refraction is at a single surface, describe two circular arcs about a common centre O (Fig. 708), the ratio of their radii being the index of refraction. Then if the incidence is from rare to dense, draw a radius O A of the smaller circle to represent the direction of the incident ray, and let N A B be the direction of the normal to the surface at the point of incidence, so that O A N is the angle of incidence. Join O B. Then O B N is the angle of refraction, since $\frac{\sin OAN}{\sin OBN} = \frac{OB}{OA}$ = index of refraction; hence O B is parallel to the refracted ray. If the incidence is from dense to rare, we must draw O B to represent the incident ray, make O B N equal to the angle of incidence, and join O A. In either case the angle A O B is the deviation, and it evidently increases with the angle of incidence O A N, attaining its greatest value when this angle (O A N″ in the figure) is a right angle, in which case the angle of refraction O B″ N″ is the critical angle.

2. To find the deviation in refraction through a prism, describe two concentric circular arcs as before (Fig. 709), the ratio of their radii being the index of refraction. Draw the radius O A of the smaller circle to represent the incident ray, N B to represent the

normal at the first surface, B N′ the normal at the second surface. Then O B represents the direction of the ray in the prism, O A′ the direction of the emergent ray, and A O A′ is accordingly the total deviation.

In fact we have

O A N = angle of incidence at first surface.
O B N = „ refraction „
O B N′ = „ incidence at second surface.
O A′ N′ = „ refraction „
A O B = deviation at first surface.
B O A′ = „ second „
A B A′ = angle between normals = angle of prism.

Again, the deviation A O A′, being the angle at the centre of a circle, is measured by the arc A A′, which subtends it. To obtain the minimum deviation, we must so arrange matters that the angle A B A′ being given (=angle of prism), the arc A A′ shall be a minimum. Let A B A′, a B a' (Fig. 710), be two consecutive positions,

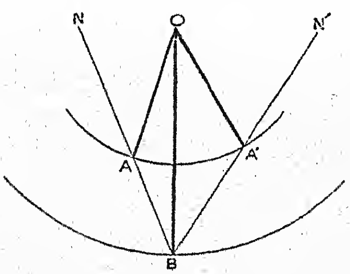


Fig. 709.—Application to Prism.

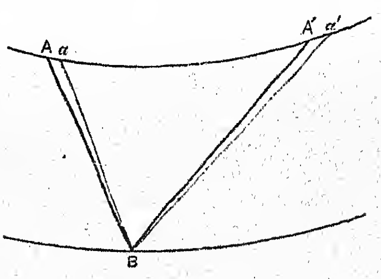


Fig. 710.—Proof of Minimum Deviation.

B A′ and B a' being greater than B A and B a. Then, since the small angles A B a, A′ B a' are equal, it is obvious, for a double reason, that the small arc A′ a' is greater than A a, and hence the whole arc a a' is greater than A A′. The deviation is therefore increased by altering the position in such a way as to make B A and B A′ depart further from equality, and is a minimum when they are equal.

997. Conjugate Foci for Minimum Deviation.—When the angle of incidence is nearly that corresponding to minimum deviation, a small change in this angle has no sensible effect on the amount of deviation.

Hence a small pencil of rays sent in this direction from a luminous point, and incident near the refracting edge, will emerge with their divergence sensibly unaltered, so that if produced backwards they

would meet in a virtual focus at the same distance (but of course not in the same direction) as the point from which they came.

In like manner, if a small pencil of rays converging towards a point, are turned aside by interposing the edge of a prism in the position of minimum deviation, they will on emergence converge to another point at the same distance. We may therefore assert that, neglecting the thickness of a prism, *conjugate foci are at the same distance from it, and on the same side, when the deviation is a minimum.*

998. Double Refraction.—Thus far we have been treating of what is called *single refraction.* We have assumed that to each given incident ray there corresponds only one refracted ray. This is true when the refraction is into a liquid, or into well-annealed glass, or into a crystal belonging to the cubic system.

On the other hand, when an incident ray is refracted into a crystal of any other than the cubic system, or into glass which is unequally stretched or compressed in different directions; for example, into unannealed glass, it gives rise in general to two refracted rays which take different paths; and this phenomenon is called *double refraction.* Attention was first called to it in 1670 by Bartholin, who observed it in the case of Iceland-spar, and its laws for this substance were accurately determined by Huygens.

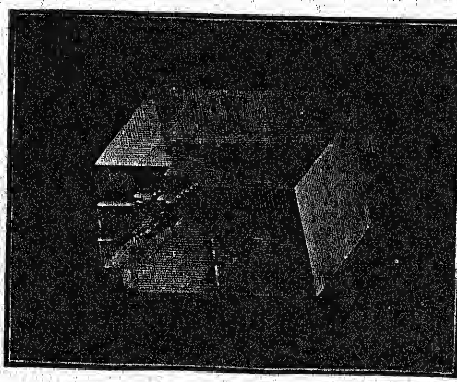


Fig. 711.—Iceland-spar.

999. Phenomena of Double Refraction in Iceland-spar.—Iceland-spar or calc-spar is a form of crystallized carbonate of lime, and is found in large quantity in the country from which it derives its name. It is usually found in rhombohedral form, as represented in Figs. 711, 712.

To observe the phenomenon of double refraction, a piece of the spar may be laid on a page of a printed book. All the letters seen through it will appear double, as in Fig. 712; and the depth of their blackness is considerably less than that of the originals, except where the two images overlap.

In order to state the laws of the phenomena with precision, it is necessary to attend to the crystalline form of Iceland-spar.

At the corner which is represented as next us in Fig. 711 three equal obtuse angles meet; and this is also the case at the opposite

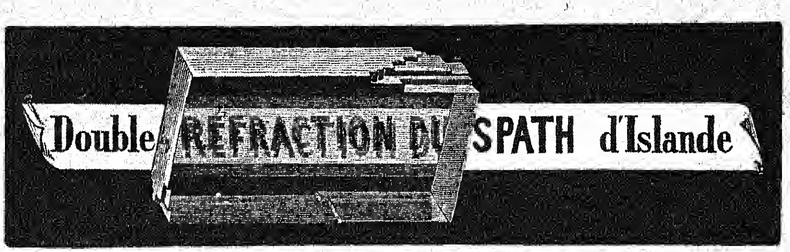


Fig. 712.—Double Refraction of Iceland-spar.

corner which is out of sight. If a line be drawn through one of these corners, making equal angles with the three edges which meet there, it or any line parallel to it is called the *axis* of the crystal; the axis being properly speaking not a definite *line* but a definite *direction.*

The angles of the crystal are the same in all specimens; but the

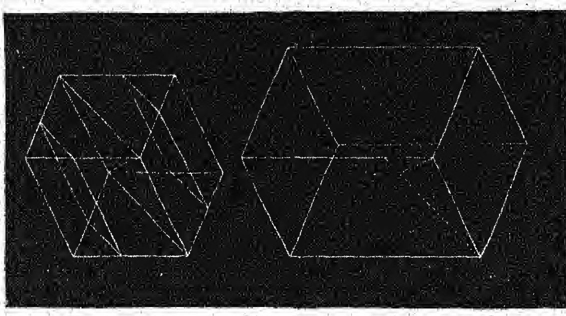


Fig. 713.—Axis of the Crystal.

lengths of the three edges (which may be called the oblique length, breadth, and thickness) may have any ratios whatever. If the crystal is of such proportions that these three edges are equal, as in the first part of Fig. 713, the axis is the direction of one of its diagonals, which is represented in the figure.

Any plane containing (or parallel to) the axis is called a *principal plane* of the crystal.

If the crystal is laid over a dot on a sheet of paper, and is made

to rotate while remaining always in contact with the paper, it will be observed that, of the two images of the dot, one remains unmoved, and the other revolves round it. The former is called the *ordinary*, and the latter the *extraordinary* image. It will also be observed that the former appears nearer than the latter, being more lifted up by refraction.

The rays which form the ordinary image follow the ordinary law of sines (§ 983). They are called the ordinary rays. Those which form the extraordinary image (called the extraordinary rays) do not follow the law of sines, except when the refracting surface is parallel to the axis, and the plane of incidence perpendicular to the axis; and in this case their index of refraction (called the extraordinary index) is different from that of the ordinary rays. The ordinary index is 1·65, and the extraordinary 1·48.

When the plane of incidence is parallel to the axis, the extraordinary ray always lies in this plane, whatever be the direction of the refracting surface; but the ratio of the sines of the angles of incidence and refraction is variable.

When the plane of incidence is oblique to the axis, the extraordinary ray generally lies in a different plane.

We shall recur to the subject of double refraction in the concluding chapter of this volume.

CHAPTER LXX.

LENSES.

1000. Forms of Lenses.—A lens is usually a piece of glass bounded by two surfaces which are portions of spheres. There are two principal classes of lenses.

1. *Converging* lenses or *convex* lenses, which have one or other of the three forms represented in Fig. 714. The first of these is called double convex, the second plano-convex, and the third concavo-convex. This last is also called a converging meniscus. All three

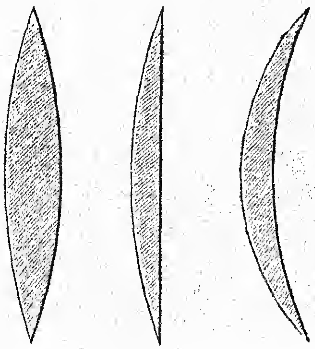


Fig. 714.—Converging Lenses.

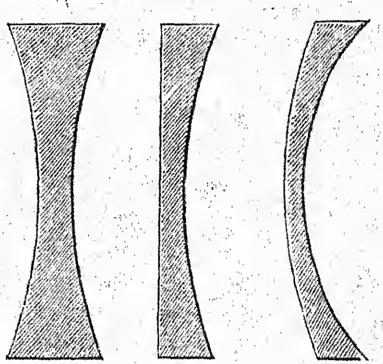


Fig. 715.—Diverging Lenses.

are thicker in the middle than at the edges. They are called converging, because rays are always more convergent or less divergent after passing through them than before.

2. *Diverging* lenses or *concave* lenses (Fig. 715) produce the opposite effect, and are characterized by being thinner in the middle than at the edges. Of the three forms represented, the first is double concave, the second plano-concave, and the third convexo-concave (also called a diverging meniscus).

From the immense importance of lenses, especially convex lenses, in practical optics, it will be necessary to explain their properties at some length.

1001. **Principal Focus.**—A lens is usually a solid of revolution, and the axis of revolution is called the *axis* of the lens, or sometimes the *principal axis.* When the surfaces are spherical, it is the line joining their centres of curvature.

When rays which were originally parallel to the principal axis pass through a convex lens (Fig. 716), the effect of the two refractions which they undergo, one on entering and the other on leaving the lens, is to make them all converge approximately to one point F, which is called the *principal focus.* The distance A F of the principal focus from the lens is called the *principal focal distance,* or more briefly and usually, the *focal length* of the lens. There is another principal focus at the same distance on the other side of the lens, corresponding to an incident beam coming in the opposite direction. The focal length depends on the convexity of the surfaces of the lens, and also on the refractive power of the material of which it is composed, being shortened either by an increase of refractive power or by a diminution of the radii of curvature of the faces.

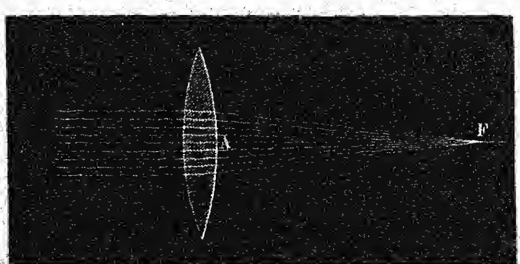


Fig. 716.—Principal Focus of Convex Lens.

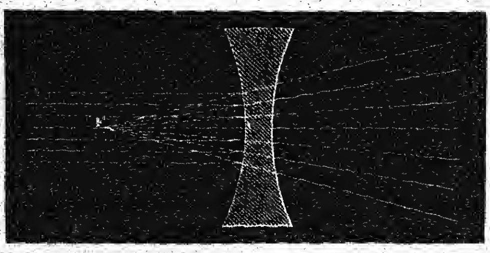


Fig. 717.—Principal Focus of Concave Lens.

In the case of a concave lens, rays incident parallel to the principal axis diverge after passing through; and their directions, if produced backwards, would approximately meet in a point F (Fig. 717), which is still called the principal focus. It is only a virtual focus, inasmuch as the emergent rays do not actually pass through it, whereas the principal focus of a converging lens is real.

1002. Optical Centre of a Lens. Secondary Axes.—Let O and O′ (Fig. 718) be the centres of the two spherical surfaces of a lens. Draw any two parallel radii O I, O′ E to meet these surfaces, and let the joining line I E represent a ray passing through the lens. This ray makes equal angles with the normals at I and E, since these latter are parallel by construction; hence the incident and emergent rays S I, E R also make equal angles with the normals, and are therefore parallel. In fact, if tangent planes (indicated by the dotted lines in the figure) are drawn at I and E, the whole course of the ray SIER will be the same as if it had passed through a plate bounded by these planes.

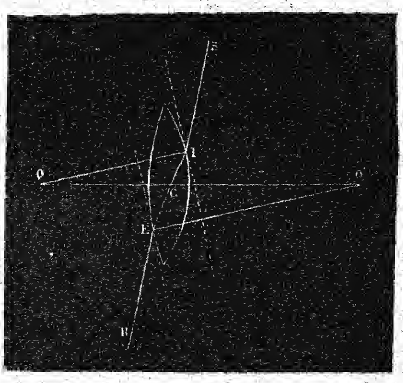


Fig. 718.—Centre of Lens.

Let C be the point in which the line I E cuts the principal axis, and let R, R′ denote the radii of the two spherical surfaces. Then, from the similarity of the triangles O C I, O′ C E, we have

$$\frac{OC}{CO'} = \frac{R}{R'}; \qquad (1)$$

which shows that the point C divides the line of centres O O′ in a definite ratio depending only on the radii. Every ray whose direction on emergence is parallel to its direction before entering the lens, must pass through the point C in traversing the lens; and conversely, every ray which, in its course through the lens, traverses the point C, has parallel directions at incidence and emergence. The point C which possesses this remarkable property is called the *centre*, or *optical centre*, of the lens.

In the case of a double convex or double concave lens, the optical centre lies in the interior, its distances from the two surfaces being directly as their radii. In plano-convex and plano-concave lenses it is situated on the convex or concave surface. In a meniscus of either kind it lies outside the lens altogether, its distances from the surfaces being still in the direct ratio of their radii of curvature.[1]

[1] These consequences follow at once from equation (1); for the distances of C from the

In elementary optics it is usual to neglect the thickness of the lens. The incident and emergent rays S I, E R may then be regarded as lying in one straight line which passes through C, and we may lay down the proposition that *rays which pass through the centre of a lens undergo no deviation.* Any straight line through the centre of a lens is called a *secondary axis.*

The approximate convergence of the refracted rays to a point, when the incident rays are parallel, is true for all directions of in-

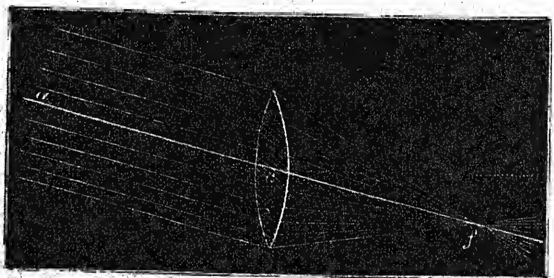


Fig. 719.—Principal Focus on Secondary Axis.

cidence; and the point to which the emergent rays approximately converge (f, Fig. 719) is always situated on the secondary axis (acf) parallel to the incident rays. The focal distance is sensibly the same as for rays parallel to the principal axis, unless the obliquity is considerable.

1003. Conjugate Foci.—When a luminous point S sends rays to a

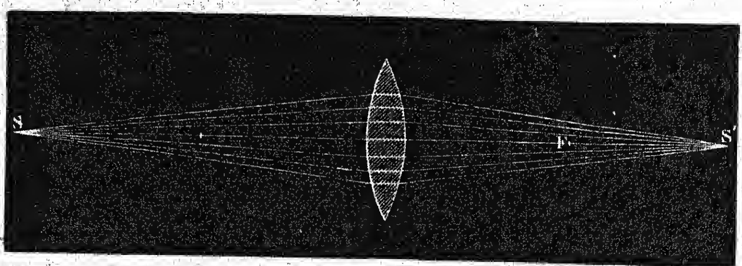


Fig. 720.—Conjugate Foci, both Real.

lens (Fig. 720), the emergent rays converge (approximately) to one

two faces are respectively the difference between R and O C, and the difference between R' and O' C, and we have

$$\frac{R}{R'} = \frac{OC}{O'C} = \frac{R - OC}{R' - O'C}.$$

point S′; whence it follows that rays sent from S′ to the lens would converge (approximately) to S. Two points thus related are called *conjugate foci* of the lens, and the line joining them always passes through the centre of the lens; in other words, they must either be both on the principal axis, or both on the same secondary axis.

The fact that rays which come from one point go to one point is the foundation of the theory of images, as we have already explained in connection with mirrors (§ 967).

The diameters of object and image are directly as their distances from the centre of the lens, and the image will be erect or inverted according as the object and image lie on the same side or on opposite sides of this centre (§ 971). There is also, in the case of lenses, the same difference between an image seen in mid-air and an image thrown on a screen which we have pointed out in § 974.

It is to be remarked that the distinction between principal and secondary axes has much more significance in the case of lenses than of mirrors; and images produced by a lens are more distinct in the neighbourhood of the principal axis than at a distance from it.

1004. Formulæ relating to Lenses.—The deviation produced in a ray by transmission through a lens will not be altered by substituting

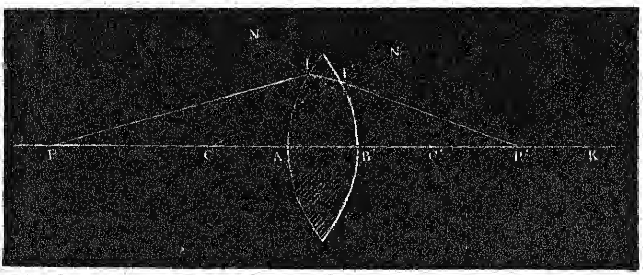


Fig. 721.—Diagram showing Path of Ray, and Normals.

for the lens a prism bounded by planes which touch the lens at the points of incidence and emergence; and in the actual use of lenses, the direction of the rays with respect to the supposed prism is such as to give a deviation not differing much from the minimum. The expression for the minimum deviation (§ 995) is $2\,i - 2\,r$ or $2\,i - A$; and when the angle of the prism is small, as it is in the case of ordinary lenses, we may assume $\frac{i}{r} = \frac{\sin i}{\sin r} = \mu$; so that $2\,i$ becomes $2\,\mu r$ or μA, and the expression for the deviation becomes

$$(\mu - 1)\,A, \qquad (1)$$

A being the angle between the tangent planes (or between the normals) at the points of entrance and emergence.

Let x_1 and x_2 denote the distances of these points respectively from the principal axis, and r_1, r_2 the radii of curvature of the faces on which they lie. Then $\frac{x_1}{r_1}, \frac{x_2}{r_2}$ are the sines of the angles which the normals make with the axis, and the angle A is the sum or difference of these two angles, according to the shape of the lens. In the case of a double convex lens it is their sum, and if we identify the sines of these small angles with the angles themselves, we have

$$A = \frac{x_1}{r_1} + \frac{x_2}{r_2}. \tag{2}$$

But if p_1, p_2 denote the distances from the faces of the lens to the points where the incident and emergent rays cut the principal axis, $\frac{x_1}{p_1}, \frac{x_2}{p_2}$ are the sines of the angles which these rays make with the axis, and the deviation is the sum or difference of these two angles, according as the conjugate foci are on opposite sides or on the same side of the lens. In the former case, identifying the angles with their sines, the deviation is $\frac{x_1}{p_1} + \frac{x_2}{p_2}$, and this, by formula (1), is to be equal to $(\mu - 1)$ A, that is, to $(\mu - 1)\left(\frac{x_1}{r_1} + \frac{x_2}{r_2}\right)$.

If the thickness of the lens is negligible in comparison with p_1, p_2, we may regard x_1 and x_2 as equal, and the equation

$$\frac{x_1}{p_1} + \frac{x_2}{p_2} = (\mu - 1)\left(\frac{x_1}{r_1} + \frac{x_2}{r_2}\right) \tag{3}$$

will reduce to

$$\frac{1}{p_1} + \frac{1}{p_2} = (\mu - 1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right). \tag{4}$$

If p_1 is infinite, the incident rays are parallel, and p_2 is the principal focal length, which we shall denote by f. We have therefore

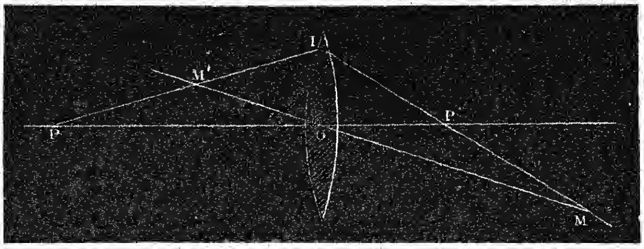
$$\frac{1}{f} = (\mu - 1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \tag{5}$$

and

$$\frac{1}{p_1} + \frac{1}{p_2} = \frac{1}{f}. \tag{6}$$

1005. Conjugate Foci on Secondary Axis.—Let M (Fig. 722) be a luminous point on the secondary axis M O M′, O being the centre of

the lens, and let M′ be the point in which an emergent ray corresponding to the incident ray M I cuts this axis. Let x denote x_1 or x_2, the distances of the points of incidence and emergence from the



Obliq. 1′ 1″

Fig. 722.—Conjugate Foci on Secondary Axis.

principal axis, and θ the obliquity of the secondary axis; then $x \cos\theta$ is the length of the perpendicular from I upon M M′, and $\frac{x\cos\theta}{\text{M I}}$, $\frac{x\cos\theta}{\text{M}'\text{I}}$, are the sines of the angles O M I, O M′ I respectively. But the deviation is the sum of these angles; hence, proceeding as in last section, we have

$$\frac{x\cos\theta}{\text{M I}}+\frac{x\cos\theta}{\text{M}'\text{I}}=(\mu-1)\left(\frac{x}{r_1}+\frac{x}{r_2}\right)=\frac{x}{f} \quad (7)$$

$$\frac{1}{\text{M I}}+\frac{1}{\text{M}'\text{I}}=\frac{1}{f\cos\theta}. \quad (8)$$

The fact that x does not appear in equations (6) and (8) shows that, for every position of a luminous point, there is a conjugate focus, lying on the same axis as the luminous point itself. Equation (8) shows that the effective focal length becomes shorter as the obliquity becomes greater, its value being $f\cos\theta$, where θ is the obliquity.

If we take account of the fact that the rays of an oblique pencil make the angles of incidence and emergence more unequal than the rays of a direct pencil and thus (by the laws of prisms) undergo larger deviation, we obtain a still further shortening of the effective focal length for oblique pencils.

When the obliquity is small, $\cos\theta$ may be regarded as unity, and we may employ the formula

$$\frac{1}{p_1}+\frac{1}{p_2}=\frac{1}{f} \quad (6)$$

for oblique as well as for direct pencils.

1006. Discussion of the Formula for Convex Lenses.—For convex

lenses f is to be regarded as positive; p will be positive when measured from the lens towards the incident light, and p' when measured in the direction of the emergent light.

Formula (6), being identical with equation (b) of § 968, leads to results analogous to those already deduced for concave mirrors.

As one focus advances from infinite distance to a principal focus, its conjugate moves away from the other principal focus to infinite distance on the other side. The more distant focus is always moving more rapidly than the nearer, and the least distance between them is accordingly attained when they are equidistant from the lens; in which case the distance of each of them from the lens is $2f$, and their distance from each other $4f$.

If either of the distances, as p, is less than f, the formula shows that the other distance p' is negative. The meaning is that the two foci are on the same side of the lens, and in this case one of them (the more distant of the two) must be virtual. For example, in Fig. 723, if S, S′ are a pair of conjugate foci, one of them S being between the principal focus F and the lens, rays sent to the lens by a luminous point at S, will, after emergence, diverge as if from S′; and rays coming from the other side of the lens, if they converge to S′ before incidence, will in reality be made to meet in S. As S moves towards the lens, S′ moves in the same direction more rapidly; and they become coincident at the surface of the lens. The formula in fact shows that if $\frac{1}{p}$ is very great in comparison with $\frac{1}{f}$, and positive, $\frac{1}{p'}$ must be very great and negative; that is to say, if p is a very small positive quantity, p' is a very small negative quantity.

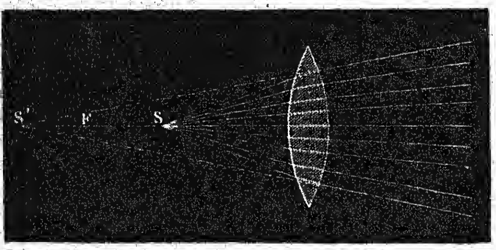


Fig. 723.—Conjugate Foci, one Real, one Virtual.

1007. Formation of Real Images.—Let A B (Fig. 724) be an object in front of a lens, at a distance exceeding the principal focal length. It will have a real image on the other side of the lens. To determine the position of the image by construction, draw through any point A of the object a line parallel to the principal axis, meeting

the lens in A′. The ray represented by this line will after refraction, pass through the principal focus F; and its intersection with the secondary axis A O determines the position of a, the focus conjugate to A. We can in like manner determine the position of b, the focus conjugate to B, another point of the object; and the joining line $a\,b$ will then be the image of the line A B. It is evident that if $a\,b$ were the object, A B would be the image.

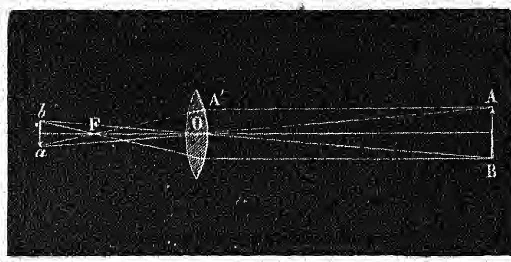


Fig. 724.—Real and Diminished Image.

Figs. 724, 725 represent the cases in which the distance of the object is respectively greater and less than twice the focal length of the lens.

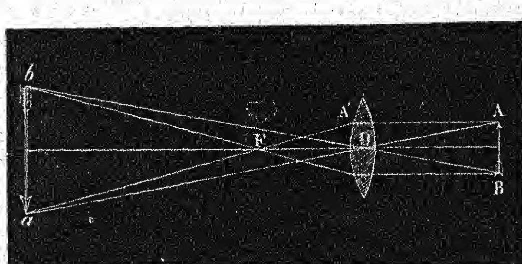


Fig. 725.—Real and Magnified Image.

1008. Size of Image.—In each case it is evident that $\frac{A\,B}{a\,b}=\frac{O\,A}{O\,a}=\frac{p}{p'}$, or *the linear dimensions of object and image are directly as their distances from the centre of the lens.*

Again, since by equation (6)

$$\frac{1}{p'}=\frac{1}{f}-\frac{1}{p}=\frac{p-f}{pf} \tag{9}$$

we have

$$\frac{p}{p'}=\frac{p-f}{f}$$

and

$$a\,b=\frac{f}{p-f}\,A\,B; \tag{10}$$

from which formula the size of the image can be calculated without finding its position.

1009. Example.—A straight line $25^{\text{mm.}}$ long is placed perpendi-

cularly on the axis, at a distance of 35 centimetres from a lens of 15 centimetres' focal length; what are the position and magnitude of the image?

To determine the distance p' we have

$$\frac{1}{35}+\frac{1}{p'}=\frac{1}{15}; \text{ whence } p'=\frac{35\times 15}{35-15}=26\tfrac{1}{4}^{\text{cm.}}$$

For the length of the image we have

$$25\frac{f}{p-f}=25\frac{15}{35-15}=18\tfrac{3}{4}^{\text{mm.}}$$

1010. Image on Cross-wires.—The position of a real image seen in mid-air can be tested by means of a cross of threads, or other convenient mark, so arranged that it can be fixed at any required point. The observer must fix this cross so that it appears approximately to coincide with a selected point of the image. He must then try whether any relative displacement of the two occurs on shifting his eye to one side. If so, the cross must be pushed nearer to the lens, or drawn back, according to the nature of the observed displacement, which follows the general rule of parallactic displacement, that the more distant object is displaced in the same direction as the observer's eye. The cross may thus be brought into exact coincidence with the selected point of the image, so as to remain in apparent coincidence with it from all possible points of view. When this coincidence has been attained, the cross is at the focus conjugate to that which is occupied by the selected point of the object.

By employing two crosses of threads, one to serve as object, and the other to mark the position of the image, it is easy to verify the fact that when the second cross coincides with the image of the first, the first also coincides with the image of the second.

1011. Aberration of Lenses.—In the investigations of §§ 1004, 1005, we made several assumptions which were only approximately true. The rays which proceed from a luminous point to a lens are in fact not accurately refracted to one point, but touch a curved surface called a caustic. The cusp of this caustic is the conjugate focus, and is the point at which the greatest concentration of light occurs. It is accordingly the place where a screen must be set to obtain the brightest and most distinct image. Rays from the central parts of the lens pass very nearly through it; but rays from the circumferen-

tial portions fall short of it. This departure from exact concurrence is called *aberration*. The distinctness of an image on a screen is improved by employing an angular diaphragm to cut off all except the central rays; but the brightness is of course diminished.

By holding a convex lens in a position very oblique to the incident light, a primary and secondary focal line can be exhibited on a screen perpendicular to the beam, just as in the case of concave mirrors (§ 975). The experiment, however, is rather more difficult of performance.

1012. Virtual Images.—Let an object A B be placed between a convex lens and its principal focus. Then the foci conjugate to the points A, B are virtual, and their positions can be found by construction from the consideration that rays through A, B, parallel to the principal axis, will be refracted to F, the principal focus on the other side. These refracted rays, if produced backward, must meet the secondary axes O A, O B in the required points. An eye placed on the other side of the lens will accordingly see a virtual image, erect, magnified, and at a greater distance from the lens than the object. This is the principle of the simple microscope. The formula for the distances D, d of object and image from the lens, when both are on the same side, is

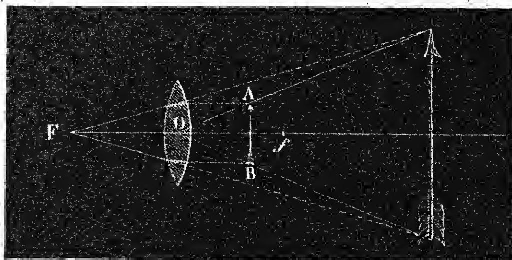


Fig. 726.—Virtual Image formed by Convex Lens.

$$\frac{1}{D} - \frac{1}{d} = \frac{1}{f}, \tag{11}$$

f denoting the principal focal length.

1013. Concave Lens.—For a concave lens, if the focal length be still regarded as positive, and denoted by f, and if the distances D, d be on the same side of the lens, the formula becomes

$$\frac{1}{d} - \frac{1}{D} = \frac{1}{f}, \tag{12}$$

which shows that d is always less than D; that is, the image is nearer to the lens than the object.

cularly on the axis, at a distance of 35 centimetres from a lens of 15 centimetres' focal length; what are the position and magnitude of the image?

To determine the distance p' we have

$$\frac{1}{35}+\frac{1}{p'}=\frac{1}{15}; \text{ whence } p'=\frac{35\times 15}{35-15}=26\tfrac{1}{4}^{\text{cm.}}$$

For the length of the image we have

$$25\frac{f}{p-f}=25\frac{15}{35-15}=18\tfrac{3}{4}^{\text{mm.}}$$

1010. Image on Cross-wires.—The position of a real image seen in mid-air can be tested by means of a cross of threads, or other convenient mark, so arranged that it can be fixed at any required point. The observer must fix this cross so that it appears approximately to coincide with a selected point of the image. He must then try whether any relative displacement of the two occurs on shifting his eye to one side. If so, the cross must be pushed nearer to the lens, or drawn back, according to the nature of the observed displacement, which follows the general rule of parallactic displacement, that the more distant object is displaced in the same direction as the observer's eye. The cross may thus be brought into exact coincidence with the selected point of the image, so as to remain in apparent coincidence with it from all possible points of view. When this coincidence has been attained, the cross is at the focus conjugate to that which is occupied by the selected point of the object.

By employing two crosses of threads, one to serve as object, and the other to mark the position of the image, it is easy to verify the fact that when the second cross coincides with the image of the first, the first also coincides with the image of the second.

1011. Aberration of Lenses.—In the investigations of §§ 1004, 1005, we made several assumptions which were only approximately true. The rays which proceed from a luminous point to a lens are in fact not accurately refracted to one point, but touch a curved surface called a caustic. The cusp of this caustic is the conjugate focus, and is the point at which the greatest concentration of light occurs. It is accordingly the place where a screen must be set to obtain the brightest and most distinct image. Rays from the central parts of the lens pass very nearly through it; but rays from the circumferen-

tial portions fall short of it. This departure from exact concurrence is called *aberration*. The distinctness of an image on a screen is improved by employing an angular diaphragm to cut off all except the central rays; but the brightness is of course diminished.

By holding a convex lens in a position very oblique to the incident light, a primary and secondary focal line can be exhibited on a screen perpendicular to the beam, just as in the case of concave mirrors (§ 975). The experiment, however, is rather more difficult of performance.

1012. Virtual Images.—Let an object A B be placed between a convex lens and its principal focus. Then the foci conjugate to the points A, B are virtual, and their positions can be found by construction from the consideration that rays through A, B, parallel to the principal axis, will be refracted to F, the principal focus on the other side. These refracted rays, if produced backward, must meet the secondary axes O A, O B in the required points. An eye placed on the other side of the lens will accordingly see a virtual image, erect, magnified, and at a greater distance from the lens than the object. This is the principle of the simple microscope. The formula for the distances D, d of object and image from the lens, when both are on the same side, is

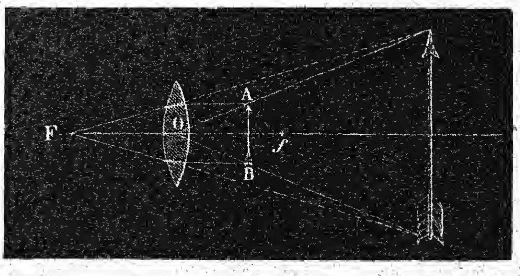


Fig. 728.—Virtual Image formed by Convex Lens.

$$\frac{1}{D} - \frac{1}{d} = \frac{1}{f}, \tag{11}$$

f denoting the principal focal length.

1013. Concave Lens.—For a concave lens, if the focal length be still regarded as positive, and denoted by f, and if the distances D, d be on the same side of the lens, the formula becomes

$$\frac{1}{d} - \frac{1}{D} = \frac{1}{f}, \tag{12}$$

which shows that d is always less than D; that is, the image is nearer to the lens than the object.

In Fig. 727, A B is the object, and $a\,b$ the image. Rays incident from A and B parallel to the principal axis will emerge as if they came from the principal focus F. Hence the points $a\,b$ are determined by the intersections of the dotted lines in the figure with the secondary axes O A, O B. An eye on the other side of the lens sees the image $a\,b$, which is always virtual, erect and diminished.

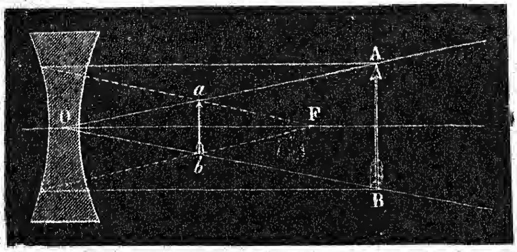


Fig. 727.—Virtual Image formed by Concave Lens.

1014. Focometer.—Silbermann's focometer (Fig. 728) is an instrument for measuring the focal lengths of convex lenses, and is based on the principle (§ 1006) that, when the object and its image are equidistant from the lens, their distance from each other is four times the focal length. It consists of a graduated rule carrying three runners M, L, M′. The middle one L is the support for the lens which is to be examined; the other two, M M′, contain two thin plates of horn or other translucent material, ruled with lines, which are at the same distance apart in both. The sliders must be adjusted until the image of one of these plates is thrown upon the other plate, without enlargement or

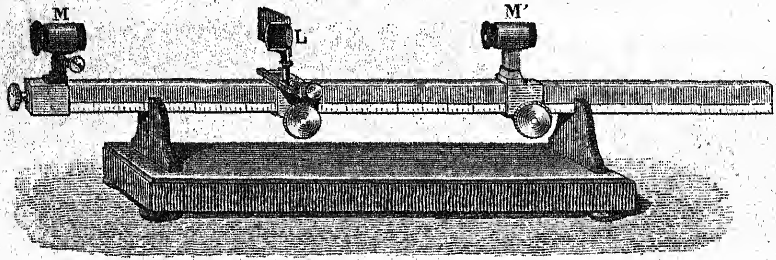


Fig. 728.—Silbermann's Focometer.

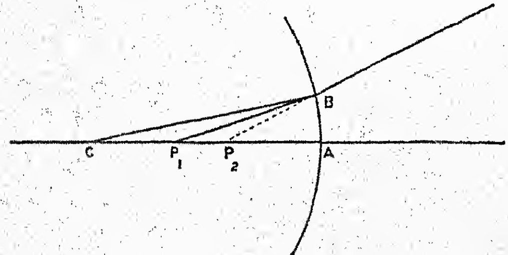


Fig. 729.

diminution, as tested by the coincidence of the ruled lines of the image with those of the plate on which it is cast. The distance between M and M′ is then read off, and divided by 4.

1015. Refraction at a Single Spherical Surface.—Suppose a small pencil of rays to be incident nearly normally upon a spherical surface which forms the boundary between two media in which the indices are μ_1 and μ_2 respectively. Let C (Fig. 729) be the centre of curvature, and C A the axis. Let P_1 be the focus of the incident, and P_2 of the refracted rays. Then for any ray $P_1 B$, $C B P_1$ is the angle of incidence and $C B P_2$ the angle of refraction. Hence by the law of sines we have (§ 993)

$$\mu_1 \sin CBP_1 = \mu_2 \sin CBP_2.$$

Dividing by sin B C A, and observing that

$$\frac{\sin CBP_1}{\sin BCA} = \frac{CP_1}{BP_1} = \frac{CP_1}{AP_1} \text{ ultimately;}$$

$$\frac{\sin CBP_2}{\sin BCA} = \frac{CP_2}{BP_2} = \frac{CP_2}{AP_2} \text{ ultimately;}$$

we obtain the equation

$$\mu_1 \frac{CP_1}{AP_1} = \mu_2 \frac{CP_2}{AP_2}, \qquad (13)$$

which expresses the fundamental relation between the positions of the conjugate foci.

Let $AC=r$, $AP_1=p_1$, $AP_2=p_2$, then equation (13) becomes

$$\mu_1 \frac{r-p_1}{p_1} = \mu_2 \frac{r-p_2}{p_2}, \qquad (14)$$

or, dividing by r,

$$\mu_1\left(\frac{1}{p_1} - \frac{1}{r}\right) = \mu_2\left(\frac{1}{p_2} - \frac{1}{r}\right),$$

which may be written

$$\frac{\mu_2}{p_2} - \frac{\mu_1}{p_1} = \frac{\mu_2 - \mu_1}{r}. \qquad (15)$$

Again, let $CA=\rho$, $CP_1=q_1$, $CP_2=q_2$, then equation (13) gives

$$\mu_1 \frac{q_1}{\rho - q_1} = \mu_2 \frac{q_2}{\rho - q_2},$$

or

$$\frac{1}{\mu_1} \frac{\rho - q_1}{q_1} = \frac{1}{\mu_2} \frac{\rho - q_2}{q_2}, \qquad (16)$$

an equation closely analogous to (14) and leading to the result (analogous to (15))

$$\frac{1}{\mu_2 q_2} - \frac{1}{\mu_1 q_1} = \left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)\frac{1}{\rho}. \qquad (17)$$

The signs of p_1, p_2, r, in (14) and (15) are to be determined by the

rule that, if one of the three points P_1, P_2, C lies on the opposite side of A from the other two, its distance from A is to be reckoned opposite in sign to theirs.

In like manner the signs of q_1, q_2, ρ, in (16) and (17) are to be determined by the rule that, if one of the three points P_1, P_2, A lies on the opposite side of C from the other two, its distance from C is to be reckoned opposite in sign to theirs.

It is usual to reckon distances positive when measured *towards the incident light;* but the formulæ will remain correct if the opposite convention be adopted.

If f denote the principal focal length, measured from A, we have, by (15), writing f for p_2 and making p_1 infinite,

$$\frac{1}{f} = \frac{\mu_2 - \mu_1}{\mu_2} \ \frac{1}{r},$$

and (15) may now be written

$$\frac{\mu_2}{p_2} - \frac{\mu_1}{p_1} = \frac{\mu_2}{f},$$

it being understood that the positive direction for f is the same as for p_1, p_2, and r.

The application of these formulæ to lenses in cases where the thickness of the lens cannot be neglected, may be illustrated by the following example.

1016. To find the position of the image formed by a spherical lens.

Let distances be measured from the centre of the sphere, and be reckoned positive on the side next the incident light.

Then, if x denote the distance of the object, y the distance of the image formed by the first refraction, z the distance of the image formed by the second refraction, a the radius of the sphere, and μ its index of refraction; we have, at the first surface,

$$\rho = a \qquad \mu_1 = 1 \qquad \mu_2 = \mu,$$

and at the second surface

$$\rho = -a \qquad \mu_1 = \mu \qquad \mu_2 = 1.$$

Hence equation (17) gives, for the first refraction,

$$\frac{1}{\mu y} - \frac{1}{x} = \left(\frac{1}{\mu} - 1\right) \frac{1}{a},$$

and for the second refraction,

$$\frac{1}{z} - \frac{1}{\mu y} = -\left(1 - \frac{1}{\mu}\right) \frac{1}{a} = \left(\frac{1}{\mu} - 1\right) \frac{1}{a}.$$

By adding these two equations, we obtain

$$\frac{1}{z} - \frac{1}{x} = \left(\frac{1}{\mu} - 1\right) \frac{2}{a} = -\frac{\mu - 1}{\mu} \cdot \frac{2}{a}.$$

If the incident rays are parallel, we have x infinite and $z = -\frac{\mu}{\mu-1}\frac{a}{2}$; that is to say, the principal focus is at a distance $\frac{\mu}{\mu-1}\frac{a}{2}$ from the centre, on the side remote from the incident light.

1017. Camera Obscura.—The images obtained by means of a hole in the shutter of a dark room (§ 938) become sharper as the size of the hole is diminished; but this diminution involves loss of light, so that it is impossible by this method to obtain an image at once bright and sharp. This difficulty can be overcome by employing a lens. If the objects in the external landscape depicted are all at distances many times greater than the focal length of the lens, their images will all be formed at sensibly the same distance from the lens, and may be received upon a screen placed at this distance. The images thus obtained are inverted, and are of the same size as if a simple aperture were employed instead of a lens. This is the principle on which the *camera obscura* is constructed.

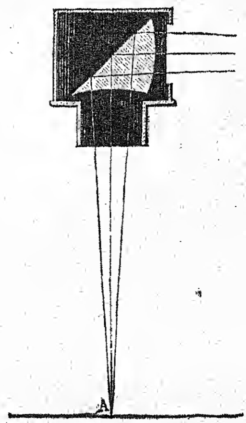


Fig. 730.—Objective of Camera.

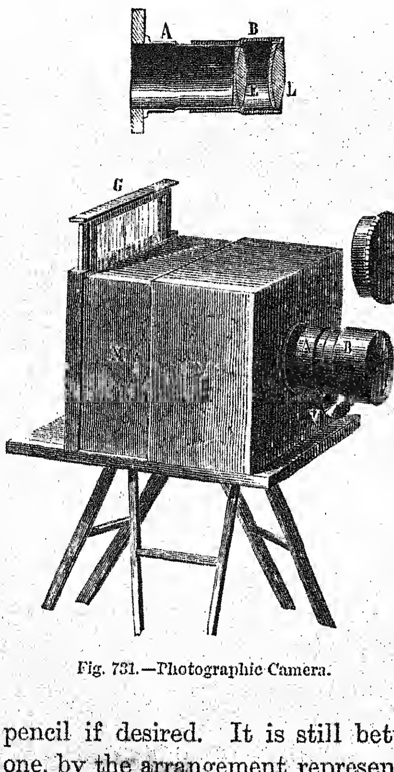


Fig. 731.—Photographic Camera.

It is a kind of tent surrounded by opaque curtains, and having at its top a revolving lantern, containing a lens with its axis horizontal, and a mirror placed behind it at a slope of 45°, to reflect the transmitted light downwards on to a sheet of white paper lying on the top of a table. Images of external objects are thus depicted on the paper, and their outlines can be traced with a pencil if desired. It is still better to combine lens and mirror in one, by the arrangement represented in section in Fig. 730. Rays

from external objects are first refracted at a convex surface, then totally reflected at the back of the lens, which is plane, and finally emerge through the bottom of the lens, which is concave, but with a larger radius of curvature than the first surface. The two refractions produce the effect of a converging mensicus. The instrument is now only employed for purposes of amusement.

1018. **Photographic Camera.**—The camera obscura employed by photographers (Fig. 731) is a box M N with a tube A B in front, containing an object-glass at its extremity. The object-glass is usually compound, consisting of two single lenses E, L, an arrangement which is very commonly adopted in optical instruments, and which has the advantage of giving the same effective focal length as a single lens of smaller radius of curvature, while it permits the employment of a larger aperture, and consequently gives more light. At G is a slide of ground glass, on which the image of the scene to be depicted is thrown, in setting the instrument. The focussing is performed in the first place by sliding the part M of the box in the part N, and finally by the pinion V which moves the lens. When the image has thus been rendered as sharp as possible, the sensitized plate is substituted for the ground glass.[1]

[1] The photographic processes at present in use are very various, both optically and chemically; but are all the same in principle with the method originally employed by Talbot. This method, which was almost forgotten during the great success of Daguerre, consists in first obtaining, on a transparent plate, a picture with lights and shades reversed, called a *negative;* then placing this upon a piece of paper sensitized with chloride of silver, and exposing it to the sun's rays. The light parts of the negative allow the light to pass and blacken the paper, thus producing a positive picture. The same negative serves for producing a great number of positives.

The negative plate is usually a glass plate covered with a film of collodion (sometimes of albumen), sensitized by a salt of silver. The following is one of the numerous formulæ for this preparation. Take

Sulphuric ether,	300	grammes
Alcohol at 40°,	200	„
Gun cotton,	5	„

Incorporate these ingredients thoroughly in a porcelain mortar; then add

Iodide of potassium,	13	grammes
Iodide of ammonium,	1·75	„
Iodide of cadmium,	1·75	„
Bromide of cadmium,	1·25	„

This mixture is poured over the plate, which is then immersed in a solution (10 per cent

1019. Use of Lenses for Purposes of Projection.—Lenses are extensively employed in the lecture-room, for rendering experiments visible to a whole audience at once, by projecting them on a screen. The arrangements vary according to the circumstances of each case, and cannot be included in a general description.

1020. Solar Microscope. Magic Lantern.—In the solar microscope, a convex lens of short focal length is employed to throw upon a screen a highly-magnified image of a small object placed a little beyond the principal focus. As the image is always much less bright than the

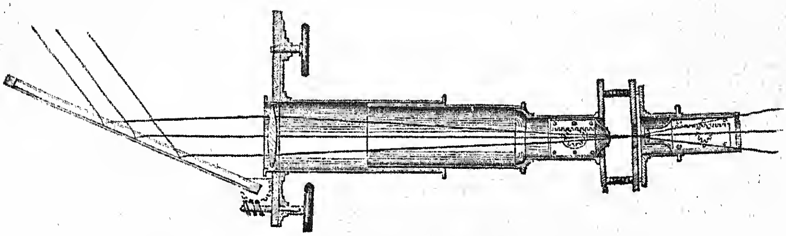


Fig. 732.—Solar Microscope.

object, and the more so as the magnification is greater, it is necessary that the object should be very highly illuminated. For this purpose the rays of the sun are directed upon it by means of a mirror and large lens; the latter serving to increase the solid angle of the cone of rays which fall upon the object, and thus to enable a larger portion of the magnifying lens to be utilized. The objects magnified are always transparent; and the images are formed by rays which have been transmitted through them.

strong) of nitrate of silver. The film of collodion is thus brought to an opal tint, and the plate, after being allowed to drain, is ready for exposure in the camera.

After being *exposed*, the picture is *developed*, by the application of a liquid for which the following is a formula;

Distilled water, ...	250	grammes.
Pyrogallic acid, ...	1	„
Crystallizable acetic acid, ...	20	„

When the picture is sufficiently developed, it is *fixed*, by the application of a solution, either of hyposulphite of soda from 25 to 30 per cent strong, or of cyanide of potassium 3 per cent strong, and the negative is completed.

To obtain a positive, the negative plate is laid upon a sheet of paper in a glass dish, the paper having been sensitized by immersing it first in a solution of common sea-salt 3 or 4 per cent strong, and then in a solution of nitrate of silver 18 per cent strong. The exposure is continued till the tone is sufficiently deep, the tint is then improved by means of a salt of gold, and the picture is fixed by hyposulphite of soda. It has then only to be washed and dried.—*D.*

The lens employed for producing the image is usually compound, consisting of a convex and a concave lens combined.

The electric light can be employed instead of the sun. The apparatus for regulating this light is usually placed within a lantern (Fig. 733), in such a position that the light is at the centre of curva-

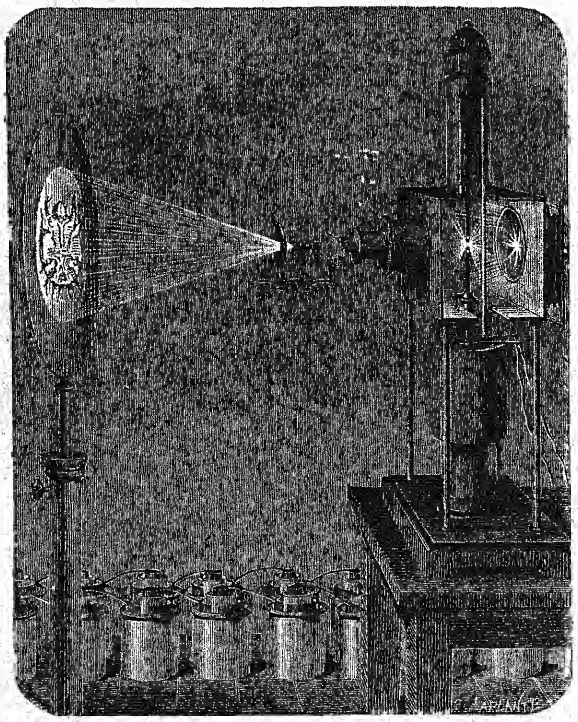


Fig. 733.—Photo-electric Microscope.

ture of a spherical mirror, so that the inverted image of the light coincides with the light itself. The light is concentrated on the object by a system of lenses, and, after passing through the object, traverses another system of lenses, placed at such a distance from the object as to throw a highly-magnified image of it on a screen. The whole arrangement is called the *electric* or *photo-electric microscope.*

The magic lantern is a rougher instrument of the same kind, employed for projecting magnified images of transparent paintings, executed on glass slides. It has one lens for converging a beam of light on the slide, and another for throwing an image of the slide on the screen. In all these cases the image is inverted.

Rev

CHAPTER LXXI.

VISION AND OPTICAL INSTRUMENTS.

1021. Description of the Eye.—The human eye (Fig. 734) is a nearly spherical ball, capable of turning in any direction in its socket. Its outermost coat is thick and horny, and is opaque except in its

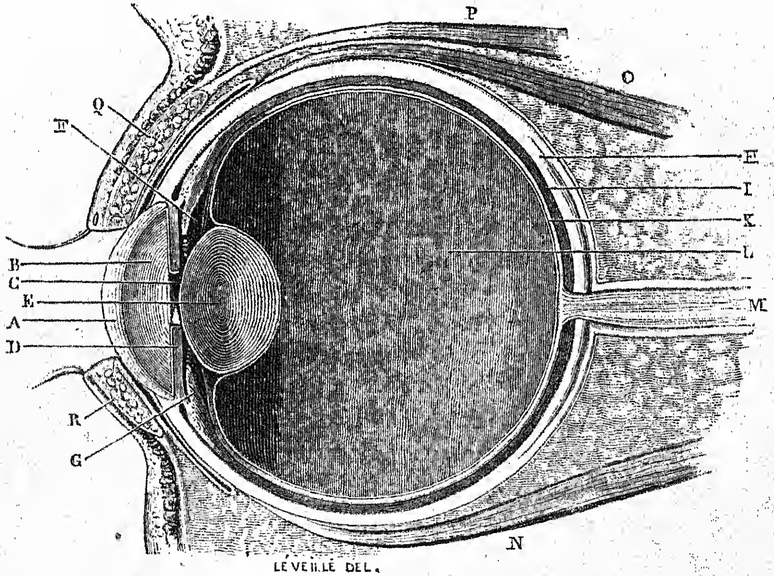


Fig. 734.—Human Eye.

anterior portion. Its opaque portion H is called the *sclerotica*, or in common language the white of the eye. Its transparent portion A is called the *cornea*, and has the shape of a very convex watch-glass. Behind the cornea is a diaphragm D, of annular form, called the *iris*. It is coloured and opaque, and the circular aperture C in its centre

outer coat Sclerotic
next Choroid Black

is called the *pupil*. By the action of the involuntary muscles of the iris, this aperture is enlarged or contracted on exposure to darkness or light. The colour of the iris is what is referred to when we speak of the colour of a person's eyes. Behind the pupil is the *crystalline lens* E, which has greater convexity at back than in front. It is built up of layers or shells, increasing in density inwards, the outermost shell having nearly the same index of refraction as the media in contact with it; an arrangement which tends to prevent the loss of light by reflection. The cavity B between the cornea and the crystalline is called the anterior chamber, and is filled with a watery liquid called the *aqueous humour*. The much larger cavity L, behind the crystalline, is called the posterior chamber, and is filled with a transparent jelly called the *vitreous humour*, inclosed in a very thin transparent membrane (the hyaloid membrane). The posterior chamber is inclosed, except in front, by the *choroid coat* or *uvea* I, which is saturated with an intensely black and opaque mucus, called the *pigmentum nigrum*. The choroid is lined, except in its anterior portion, with another membrane K, called the retina, which is traversed by a ramified system of nerve filaments diverging from the optic nerve M. Light incident on the retina gives rise to the sensation of vision; and there is no other part of the eye which possesses this property.

1022. **The Eye as an Optical Instrument.**—It is clear, from the above description, that a pencil of rays entering the eye from an external point will undergo a series of refractions, first at the anterior surface of the cornea, and afterwards in the successive layers of the crystalline lens, all tending to render them convergent (see table of indices, § 986). A real and inverted image is thus formed of any external object to which the eye is directed. If this image falls on the retina, the object is seen; and if the image thus formed on the retina is sharp and sufficiently luminous, the object is seen distinctly.

1023. **Adaptation to Different Distances.**—As the distance of an image from a lens varies with the distance of the object, it would only be possible to see objects distinctly at one particular distance, were there not special means of adaptation in the eye. Persons whose sight is not defective can see objects in good definition at all distances exceeding a certain limit. When we wish to examine the minute details of an object to the greatest advantage, we hold it at a particular distance, which varies in different individuals, and averages about eight inches. As we move it further away, we

experience rather more ease in looking at it, though the diminution of its apparent size, as measured by the visual angle, renders its minuter features less visible. On the other hand, when we bring it nearer to the eye than the distance which gives the best view, we cannot see it distinctly without more or less effort and sense of strain; and when we have brought it nearer than a certain lower limit (averaging about six inches), we find distinct vision no longer possible. In looking at very distant objects, if our vision is not defective, we have very little sense of effort. These phenomena are in accordance with the theory of lenses, which shows that when the distance of an object is a large multiple of the focal length of the lens, any further increase, even up to infinity, scarcely alters the distance of the image; but that, when the object is comparatively near, the effect of any change of its distance is considerable. There has been much discussion among physiologists as to the precise nature of the changes by which we adapt our eyes to distinct vision at different distances. Such adaptation might consist either in a change of focal length, or in a change of distance of the retina. Observations in which the eye of the patient is made to serve as a mirror, giving images by reflection at the front of the cornea, and at the front and back of the crystalline, have shown that the convexity of the front of the crystalline is materially changed as the patient adapts his eye to near or remote vision, the convexity being greatest for near vision. This increase of convexity corresponds to a shortening of focal length, and is thus consistent with theory.

1024. Binocular Vision.—The difficulty which some persons have felt in reconciling the fact of an inverted image on the retina with the perception of an object in its true position, is altogether fanciful, and arises from confused notions as to the nature of perception.

The question as to how it is that we see objects single with two eyes, rests upon a different footing, and is not to be altogether explained by habit and association.[1] To each point in the retina of one eye there is a *corresponding point*, similarly situated, in the other. An impression produced on one of these points is, in ordinary circumstances, undistinguishable from a similar impression produced on the other, and when both at once are similarly impressed, the effect is simply more intense than if one were impressed alone; or, to describe the same phenomena subjectively, we have only one field

[1] Binocular vision is a subject which has been much debated. For the account here given of it, the Editor is responsible.

of view for our two eyes, and in any part of this field of view we see either one image, brighter than we should see it by one alone, or else we see two overlapping images. This latter phenomenon can be readily illustrated by holding up a finger between one's eyes and a wall, and looking at the wall. We shall see, as it were, two transparent fingers projected on the wall. One of these transparent fingers is in fact seen by the right eye, and the other by the left, but our visual sensations do not directly inform us which of them is seen by the right eye, and which by the left.

The principal advantage of having two eyes is in the estimation of distance, and the perception of relief. In order to see a point as single by two eyes, we must make its two images fall on corresponding points of the retinæ; and this implies a greater or less convergence of the optic axes according as the object is nearer or more remote. We are thus furnished with a direct indication of the distance of the object from our eyes; and this indication is much more precise than that derived from the adjustment of their focal length.

In judging of the comparative distances of two points which lie nearly in the same direction, we are greatly aided by the parallactic displacement which occurs when we change our own position.

We can also form an estimate of the nearness of an object, from the amount of change in its apparent size, contour, and bearing, produced by shifting our position. This would seem to be the readiest means by which very young animals can distinguish near from remote objects.

1025. Stereoscope.—The perception of relief is closely connected with the doubleness of vision which occurs when the images on corresponding portions of the two retinæ are not similar. In surveying an object we run our eyes rapidly over its surface, in such a way as always to attain single vision of the particular point to which our attention is for the instant directed. We at the same time receive a somewhat indistinct impression of all the points within our field of view; an impression which, when carefully analysed, is found to involve a large amount of doubleness. These various impressions combine to give us the perception of relief; that is to say, of *form in three dimensions.*

The perception of relief in binocular vision is admirably illustrated by the *stereoscope,* an instrument which was invented by Wheatstone, and reduced to its present more convenient form by Brewster. Two figures are drawn, as in Fig. 735, being perspective representations of

the same object from two neighbouring points of view, such as might be occupied by the two eyes in looking at the object. Thus if the object be a cube, the right eye will have a fuller view of the right

Fig. 735.—Stereoscopic Pictures.

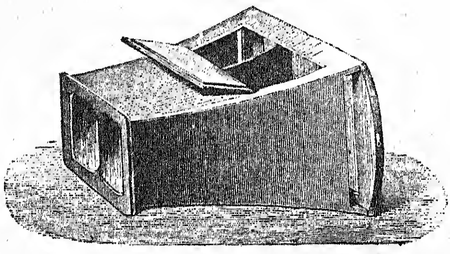


Fig. 736.—Stereoscope.

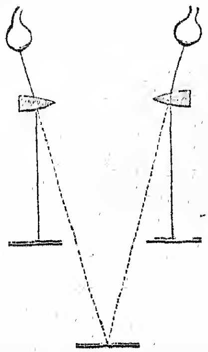


Fig. 737.—Path of Rays in Stereoscope.

face, and the left eye of the left face. The two pictures are placed in the right and left compartments of a box, which has a partition down the centre serving to insure that each eye shall see only the picture intended for it; and over each of the compartments a half-lens is fixed, serving, as in Fig. 737, not only to magnify the picture, but at the same time to displace it, so that the two virtual images are brought into approximate coincidence. Stereoscopic pictures are usually photographs obtained by means of a double camera, having two objectives, one beside the other, which play the part of two eyes.

When matters are properly arranged, the observer seems to see the object in relief. He finds himself able to obtain single view of any one point of the solid image which is before him; and the adjustments of the optic axes which he finds it necessary to make, in shifting his view from one point of it to another, are exactly such as would be required in looking at a solid object.

When one compartment of the stereoscope is empty, and the other contains an object, an observer, of normal vision, looking in in the ordinary way, is unable to say which eye sees the object. If two pictures are combined, consisting of two equal circles, one of them

having a cross in its centre, and the other not, he is unable to decide whether he sees the cross with one eye or both.

When two entirely dissimilar pictures are placed in the two compartments, they compete for mastery, each of them in turn becoming more conspicuous than the other, in spite of any efforts which the observer may make to the contrary. A similar fluctuation will be observed on looking steadily at a real object which is partially hidden from one eye by an intervening object. This tendency to alternate preponderance renders it well nigh impossible to combine two colours by placing one under each eye in the stereoscope.

The immediate visual impression, when we look either at a real solid object, or at the apparently solid object formed by properly combining a pair of stereoscopic views, is a single picture formed of two slightly different pictures superimposed upon each other. The coincidence becomes exact at any point to which attention is directed, and to which the optic axes are accordingly made to converge, but in the greater part of the combined picture there is a want of coincidence, which can easily be detected by a collateral exercise of attention. The fluctuation above described to some extent tends to conceal this doubleness; and in looking at a real solid object, the concealment is further assisted by the blurring of parts which are out of focus.

1026. Visual Angle. Magnifying Power.—The angle which a given straight line subtends at the eye is called its *visual angle*, or the *angle under which it is seen.* This angle is the measure of the length of the image of the straight line on the retina. Two discs at different distances from the eye, are said to have the same apparent size, if their diameters are seen under equal angles. This is the condition that the nearer disc, if interposed between the eye and the remoter disc, should be just large enough to conceal it from view.

The angle under which a given line is seen, evidently depends not only on its real length, and the direction in which it points, but also on its distance from the eye; and varies, in the case of small visual angles, in the inverse ratio of this distance. The *apparent length* of a straight line may be regarded as measured by the visual angle which it subtends.

By the *magnifying power* of an optical instrument, is usually meant the ratio in which it increases *apparent lengths* in this sense. In the case of telescopes, the comparison is between an object as

seen in the telescope, and the same object as seen with the naked eye at its actual distance. In the case of microscopes, the comparison is between the object as seen in the instrument, and the same object as seen by the naked eye at the least distance of distinct vision, which is usually assumed as 10 inches.

But two discs, whose diameters subtend the same angle at the eye, may be said to have the same *apparent area;* and since the areas of similar figures are as the squares of their linear dimensions, it is evident that the apparent area of an object varies as the square of the visual angle subtended by its diameter. The number expressing *magnification of apparent area* is therefore the square of the magnifying power as above defined. Frequently, in order to show that the comparison is not between apparent areas, but between apparent lengths, an instrument is said to magnify so many *diameters*. If the diameter of a sphere subtends 1° as seen by the naked eye, and 10° as seen in a telescope, the telescope is said to have a magnifying power of 10 diameters. The superficial magnification in this case is evidently 100.

The apparent length and apparent area of an object are respectively proportional to the length and area of its image on the retina.

Apparent length is measured by the plane angle, and apparent area by the solid angle, which an object subtends at the eye.

1027. Spectacles.—Spectacles are of two kinds, intended to remedy two opposite defects of vision. Short-sighted persons can see objects distinctly at a smaller distance than persons whose vision is normal; but always see distant objects confused. On the other hand, persons whose vision is normal in their youth, usually become over-sighted with advancing years, so that, while they can still adjust their eyes correctly for distant vision, objects as near as 10 or 12 inches always appear blurred. Spectacles for over-sighted persons are convex, and should be of such focal length, that, when an object is held at about 10 inches distance, its virtual image is formed at the nearest distance of distinct vision for the person who is to use them. This latter distance must be ascertained by trial. Call it p inches; then, by § 1012, the formula for computing the required focal length x (in inches) is

$$\frac{1}{10} - \frac{1}{p} = \frac{1}{x}.$$

For example, if 15 inches is the nearest distance at which the person

can conveniently read without spectacles, the focal length required is 30 inches.

In Fig. 738, A represents the position of a small object, and A′ that of its virtual image as seen with spectacles of this kind.

Over-sight is not the only defect which the eye is liable to acquire

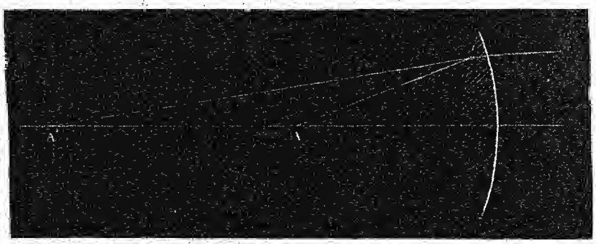


Fig. 738.—Spectacle-glass for Over-sighted Eye.

by age, but it is the defect which ordinary spectacles are designed to remedy.

Spectacles for short-sighted persons are concave, and the focal

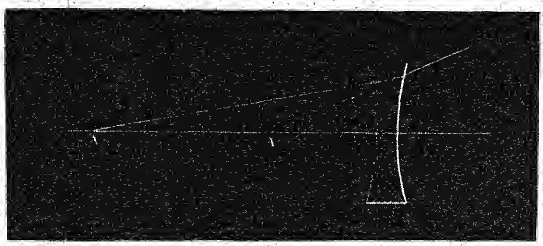


Fig. 739.—Spectacle-glass for Short-sighted Eye.

length which they ought to have, if designed for reading, may be computed by the formula

$$\frac{1}{p} - \frac{1}{10} = \frac{1}{x},$$

p denoting the nearest distance at which the person can read, and x the focal length, both in inches. If his *greatest* distance of distinct vision exceeds the focal length, he will be able, by means of the spectacles, to obtain distinct vision of objects at all distances, from 10 inches upwards.

1028. Simple Magnifier.—A *magnifying glass* is a convex lens, of

astigmatism

shorter focal length than the human eye, and is placed at a distance somewhat less than its focal length from the object to be viewed. In Fig. 740, ab is the object, and A B the virtual image which is seen by the eye K. The construction which we have employed for drawing the image is one which we have several times used before.

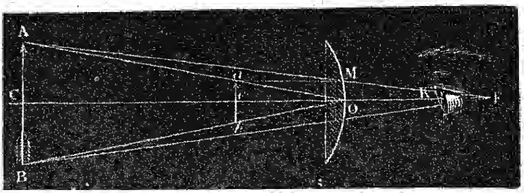


Fig. 740.—Magnifying Glass.

Through the point a, the line a M is drawn parallel to the principal axis. F M is then drawn from the principal focus F; O a is drawn from the optical centre O; and these two lines are produced till they meet in A.

Distance of lens from object. In order that the image may be properly seen, its distance from the eye must fall between the limits of distinct vision; and in order that it may be seen under the largest possible visual angle, the eye must be close to the lens, and the object must be as near as is compatible with distinct vision. This and other interesting properties are established by the following investigation:—

Let θ denote the visual angle under which the observer sees the image of the portion ac of the object. Also let x denote the distance cO of the object from the lens, and y the distance O K of the lens from the eye. Then we have

$$\tan\theta = \frac{\text{A C}}{\text{CK}} = \frac{\text{A C}}{\text{C O}+y};$$

but, by formulæ (10) and (11) of last chapter, we have

$$\text{A C} = ac\frac{f}{f-x}, \quad \text{C O} = x.\frac{f}{f-x}.$$

Substituting these values for A C and C O, and reducing, we have

$$\tan\theta = ac.\frac{f}{(x+y)f-xy}. \qquad \text{(A)}$$

This equation shows that, for a given lens and a given object, the visual angle varies inversely as the quantity $(x+y)f-xy$.

The following practical consequences are easily drawn:—

(1) If the distance $x+y$ of the eye from the object is given, the visual angle increases as the two distances x, y approach equality, and is not altered by interchanging them.

(2) If one of the two distances x, y be given, and be less than f, the other must be made as small as possible, if we wish to obtain the largest possible visual angle.

To obtain the absolute maximum of visual angle, we must select, from the various positions which make C K equal to the nearest distance of distinct vision, that which gives the largest value of A C, since the quotient of A C by C K is the tangent of the visual angle. Now A C increases as the image moves further from the lens, and hence the absolute maximum is obtained by making its distance from the lens equal to the nearest distance of distinct vision, and making the eye come up close to the lens. In this case the distance p of the object from the lens is given by the equation $\frac{1}{p} - \frac{1}{\mathrm{D}} = \frac{1}{f}$, where D denotes the nearest distance of distinct vision; and tan θ is $\frac{ac}{p}$ or $ac\left(\frac{1}{f}+\frac{1}{\mathrm{D}}\right)$. But the greatest angle under which the body could be seen by the naked eye is the angle whose tangent is $\frac{ac}{\mathrm{D}}$; hence the visual angle (or its tangent) is increased by the lens in the ratio $1+\frac{\mathrm{D}}{f}$, which is called the *magnifying power*. If the object were in the principal focus, and the eye close to the lens, the magnifying power would be $\frac{\mathrm{D}}{f}$. In either case, the thickness of the lens being neglected, the visual angle is the angle which the object subtends at the centre of the lens, and therefore varies inversely as the distance of this centre from the object. For lenses of small focal length, the reciprocal of the focal length may be regarded as proportional to the magnifying power.

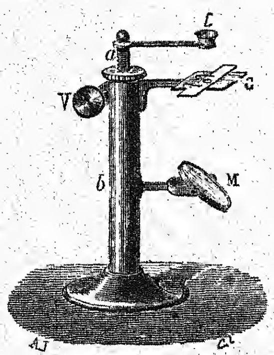


Fig. 741.—Simple Microscope.

Simple Microscope.—By a *simple microscope* is usually understood a lens of short focal length mounted in a manner convenient for the examination of small objects. Fig. 741 represents an instrument of this kind. The lens l is mounted in brass, and carried at the end of an arm. It is raised and lowered by turning the milled head V,

which acts on the rack a. C is the platform on which the object is laid, and M is a concave mirror, which can be employed for increasing the illumination of the object.

1029. Compound Microscope.—In the compound microscope, there is one lens which forms a real and greatly enlarged image of the object; and this image is itself magnified by viewing it through another lens

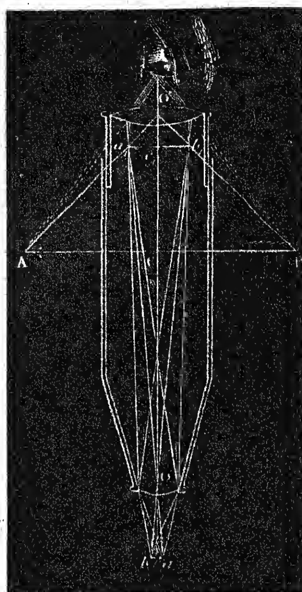


Fig. 742.—Compound Microscope.

In Fig. 742, $a\,b$ is the object, O is the first lens, called the *objective,* and is placed at a distance only slightly exceeding its focal length from the object; an inverted image $a_1\,b_1$ is thus formed at a much greater distance on the other side of the lens, and proportionally larger. O′ is the second lens, called the *ocular* or *eye-piece,* which is placed at a distance a little less than its focal length from the first image $a_1\,b_1$, and thus forms an enlarged virtual image of it A B, at a convenient distance for distinct vision.

If we suppose the final image A B to be at the least distance of distinct vision from the eye placed at O′ (this being the arrangement which gives the largest visual angle), the magnifying power will be simply the ratio of the length of this image to that of the object $a\,b$, and will be the product of the two factors $\frac{\mathrm{A\,B}}{a_1\,b_1}$ and $\frac{a_1\,b_1}{a\,b}$. The former is the magnification produced by the eye-piece, and is, as we have just shown (§ 1028), $1+\frac{\mathrm{D}}{f}$. The other factor $\frac{a_1\,b_1}{a\,b}$ is the magnification produced by the objective, and is equal to the ratio of the distances $\frac{\mathrm{O}\,a_1}{\mathrm{O}\,a}$. If the objective is taken out, and replaced by another of different focal length, the readjustment will consist in altering the distance O a, leaving the distance O a_1 unchanged. The total magnification therefore varies inversely as O a, that is, nearly in the inverse ratio of the focal length of the objective. Compound microscopes are usually provided with several objectives, of various focal lengths, from which the observer makes a selection according to the magnifying power which he requires for the object to be examined. The powers most used range from 50 to 350 diameters.

The magnifying power of a microscope can be determined by direct observation, in the following way. A plane reflector pierced with a hole in its centre, is placed directly over the eye-piece (Fig. 743), at an inclination of 45°, and another plane reflector, or still better, a totally reflecting prism, as in the figure, is placed parallel to it at the distance of an inch or two, so that the eye, looking down upon the first mirror, sees, by means of two successive reflections, the image of a divided scale placed at a distance of 8 or 10 inches below the second reflector. In taking an observation, a micrometer scale engraved on glass, its divisions being at a known distance apart (say $\frac{1}{100}$ of a millimetre), is placed in the microscope as the object to be magnified; and the observer holds his eye in such a position that, by means of different parts of his pupil, he sees at once the magnified image of the micrometer scale in the microscope, and the reflected and unmagnified image of the other scale. The two images will be superimposed in the same field of view; and it is easy to observe how many divisions of the one coincide with a given number of divisions of the other. Let the divisions on the large scale be millimetres, and those on the micrometer scale hundredths of a millimetre. Then the magnifying power is 100, if one of the magnified covers one of the unmagnified divisions; and is $\frac{100\,\mathrm{N}}{n}$, if n of the former cover N of the latter. This is on the assumption that the large scale is placed at the nearest distance of convenient vision. In stating the magnifying power of a microscope, this distance is usually reckoned as 10 inches.

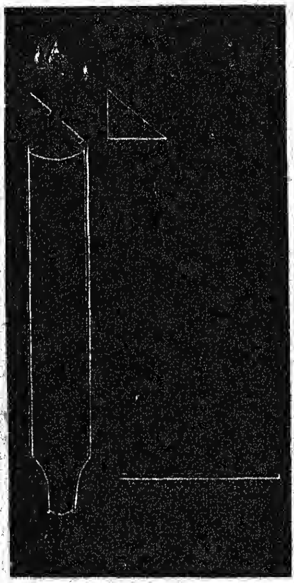


Fig. 743.
Measurement of Magnifying Power.

A short-sighted person sees an image in a microscope (whether simple or compound) under a larger visual angle than a person of normal sight; but the inequality is not so great as in the case of objects seen by the naked eye. In fact, if f be the focal length of the eye-piece in a compound microscope, or of the microscope itself if simple, and D the nearest distance of distinct vision for the

observer, the visual angle under which the image is seen in the microscope is proportional to $\frac{1}{f} + \frac{1}{D}$, the greatest visual angle for the naked eye being represented by $\frac{1}{D}$. Both these angles increase as D diminishes, but the latter increases in a greater ratio than the former. When f is as small as $\frac{1}{10}$ of an inch, the visual angle in the microscope is sensibly the same for short as for normal sight.

Before reading off the divisions in the observation above described, care should be taken to focus the microscope in such a way, that the image of the micrometer scale is at the same distance from the eye as the image of the large scale with which it is compared. When this is done, a slight motion of the eye does not displace one image with respect to the other.

Instead of a single eye-lens, it is usual to employ two lenses separated by an interval, that which is next the eye being called the *eye-glass*, and the other the *field-glass*. This combination is equivalent to the Huygenian or negative eye-piece employed in telescopes (§ 1070).

1030. Astronomical Telescope.—The astronomical refracting telescope consists essentially (like the compound microscope) of two lenses, one of which forms a real and inverted image of the object, which is looked at through the other.

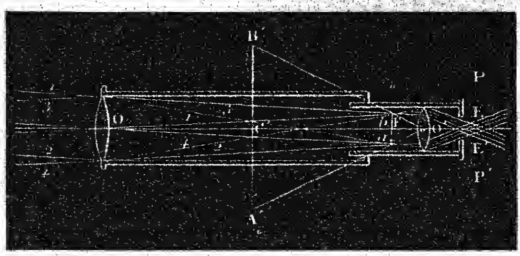


Fig. 744.—Astronomical Telescope.

In Fig. 744, O is the object-glass, which is sometimes a foot or more in diameter, and is always of much greater focal length than the eye-piece O′. The inverted image of a distant object is formed at the principal focus F. This image is represented at $a\,b$. The parallel rays marked 1, 2 come from the upper extremity of the object, and meet at a; and the parallel rays 3, 4, from the other extremity, meet at b. A′ B′ is the virtual image of $a\,b$ formed by the eye-piece. Its distance from the eye can be changed by pulling out or pushing in the eye-tube; and may in practice have any value intermediate between the least distance of distinct vision and infinity, the visual angle under which it is seen being but slightly affected by

this adjustment. The rays from the highest point of the object emerge from the eye-piece as a pencil diverging from A′; and the rays from the lowest point of the object form a pencil diverging from B′.

Magnification.—The angle under which the object would be seen by the naked eye is $a\,O\,b$; for the rays $a\,O$, $b\,O$, if produced, would pass through its extremities. The angle under which it is seen in the telescope, if the eye be close to the eye-lens, is $A'\,O'\,B'$ or $a\,O'\,b$.

The magnification is therefore $\frac{a\,O'\,b}{a\,O\,b}$, which is approximately the same as the ratio of the distances of the image ab from the two lenses $\frac{O\,F}{O'\,F}$. If the eye-tube is so adjusted as to throw the image $A'\,B'$ to infinite distance, F will be the principal focus of both lenses, and the magnification is the ratio of the focal length of the object-glass to that of the eye-piece. If the eye-tube be pushed in as far as is compatible with distinct vision (the eye being close to the lens), the magnification is greater than this in the ratio $\frac{D+f}{D}$, D denoting the nearest distance of distinct vision, and f the focal length of the eye-piece

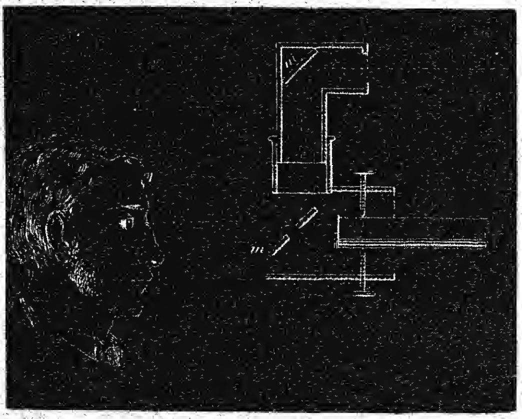


Fig. 745.—Measurement of Magnifying Power.

Who? Where

The magnification can be directly observed by looking with one eye through the telescope at a brick wall, while the other eye is kept open. The image will thus be superimposed on the actual wall, and we have only to observe how many courses of the latter coincide with a single course of the magnified image.

If the telescope is large, its tube may prevent the second eye from seeing the wall, and it may be necessary to employ a reflecting arrangement, as in Fig. 745, analogous to that described in connection with the microscope.

Telescopes without stands seldom magnify more than about 10 diameters. Powers of from 20 to 60 are common in telescopes with

stands, intended for terrestrial purposes. The powers chiefly employed in astronomical observation are from 100 to 500.

Mechanical Arrangements.—The achromatic object-glass O is set in a mounting which is screwed into one end of a strong brass tube AA (Fig. 746). In the other end slides a smaller tube F containing the eye-piece O′; and by turning the milled head V in one direction or the other, the eye-piece is moved forwards or backwards.

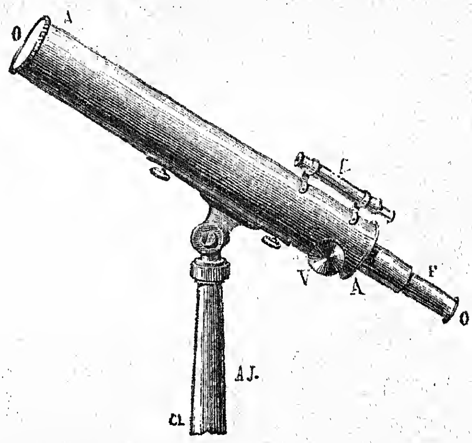


Fig. 746.—Astronomical Telescope.

Finder.—The small telescope *l*, which is attached to the principal telescope, is called a *finder*. This appendage is indispensable when the principal telescope has a high magnifying power; for a high magnifying power involves a small field of view, and consequent difficulty in directing the telescope so as to include a selected object within its range. The finder is a telescope of large field; and as it is set parallel to the principal telescope, objects will be visible in the latter if they are seen in the centre of the field of view of the former.

1031. Best Position for the Eye.—The eye-piece forms a real and inverted image of the object-glass[1] at E E′ (Fig. 744), through which all rays transmitted by the telescope must of necessity pass. If the telescope be directed to a bright sky, and a piece of white paper held behind the eye-piece to serve as a screen, a circular spot of light will be formed upon it, which will become sharply defined (and at the same time attain its smallest size) when the screen is held in the correct position. This image (which we shall call the *bright spot*) may be regarded as marking the proper place for the pupil of the observer's eye. Every ray which traverses the centre of the object-glass traverses the centre of this spot; every ray which traverses the upper edge of the object-glass traverses the lower edge of the

[1] Or it may be called an image of *the aperture which the object-glass fills*. It remains sensibly unchanged on removing the object-glass so as to leave the end of the telescope open.

spot; and any selected point of the spot receives all the rays which have been transmitted by one particular point of the object-glass. An eye with its pupil anywhere within the limits of the bright spot, will therefore see the whole field of view of the telescope. If the spot and pupil are of exactly the same size, they must be made to coincide with one another, as the necessary condition of seeing the whole field of view with the brightest possible illumination. Usually in practice the spot is much smaller than the pupil, so that these advantages can be obtained without any nicety of adjustment; but to obtain the most distinct vision, the centre of the pupil should coincide as closely as possible with the centre of the spot. To facilitate this adjustment, a brass diaphragm, with a hole in its centre, is screwed into the eye-end of the telescope, the proper place for the eye being close to this hole.

One method of determining the magnifying power of a telescope consists in measuring the diameter of the bright spot, and comparing it with the effective aperture of the object-glass. In fact, let F and f denote the focal lengths of object-glass and eye-piece, and a the distance of the spot from the centre of the eye-piece; then $F+f$ is approximately the distance of the object-glass from the same centre, and, by the formula for conjugate focal distances, we have $\frac{1}{F+f}+\frac{1}{a}=\frac{1}{f}$. Multiplying both sides of this equation by $F+f$, and then subtracting unity, we have $\frac{F+f}{a}=\frac{F}{f}$. But the ratio of the diameter of the object-glass to that of its image is $\frac{F+f}{a}$; and $\frac{F}{f}$ is the usual formula for the magnifying power. Hence, *the linear magnifying power of a telescope is the ratio of the diameter of the object-glass to that of the bright spot.*

1032. Terrestrial Telescope.—The astronomical telescope just described gives inverted images. This is no drawback in astronomical observation, but would be inconvenient in viewing terrestrial objects. In order to re-invert the image, and thus make it erect, two additional lenses O″ O‴ (Fig. 747) are introduced between the real image $a\,b$ and the eye-lens O′. If the first of these O″ is at the distance of its principal focal length from $a\,b$, the pencils which fall upon the second will be parallel, and an erect image $a'\,b'$ will thus be formed in the principal focus of O‴. This image is viewed through the eye-lens O′, and the virtual image A′ B′ which is perceived by the eye will therefore be erect. The two lenses O″, O‴, are usually

made precisely alike, in which case the two images $a\,b$, $a'\,b'$ will be equal. In the better class of terrestrial telescopes, a different ar-

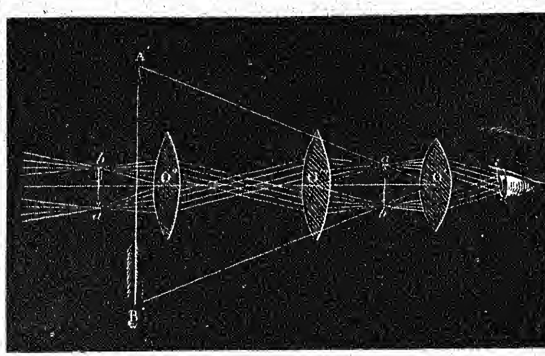


Fig. 747.—Terrestrial Eye-piece.

rangement is adopted, requiring one more lens; but whatever system be employed, the reinversion of the image always involves some loss both of light and of distinctness.

1033. Galilean Telescope.—Besides the disadvantages just mentioned, the erecting eye-piece involves a considerable addition to the length of the instrument. The telescope invented by Galileo, and the earliest of all telescopes, gives erect images with only two lenses, and with shorter length than even the astronomical telescope. O (Fig. 748) is the object-glass, which is convex as in the astronomical telescope, and would form a real and inverted image $a\,b$ at its principal focus; but the eye-glass O′, which is a concave lens, is interposed at a distance equal to or slightly exceeding its own focal length from the place of this image, and forms an erect virtual image A′ B′, which the observer sees.

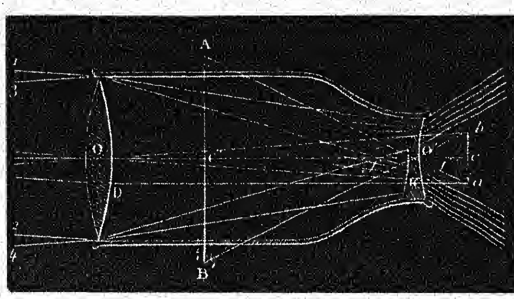


Fig. 748.—Galilean Telescope.

Neglecting the distance of his eye from the lens, the angle under which he sees the image is A′ O′ B′, which is equal to a O′ b, whereas

the visual angle to the naked eye would be $a\,O\,b$. The magnification is therefore $\frac{a\,O'\,b}{a\,O\,b}$, which is approximately equal to $\frac{O\,c}{O'\,c}$, c being the principal focus of the object-glass. If the instrument is focussed in such a way that the image A′ B′ is thrown to infinite distance, c is also the principal focus of the eye-lens, and the magnification is simply the ratio of the focal lengths of the two lenses. This is the same rule which we deduced for the astronomical telescope; but the Galilean telescope, if of the same power, is shorter by twice the focal length of the eye-lens, since the distance between the two lenses is the difference instead of the sum of their focal lengths.

This telescope has the disadvantage of not admitting of the employment of cross wires; for these, in order to serve their purpose, must coincide with the real image; and no such image exists in this telescope.

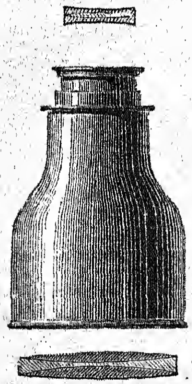


Fig. 749.—Opera-glass.

There is another peculiarity in the absence of the *bright spot* above described, the image of the object-glass formed by the eye-glass being virtual. In other telescopes, if half the object-glass be covered, half the bright spot will be obliterated; but the remaining half suffices for giving the whole field of view, though with diminished brightness. In the Galilean telescope, on the contrary, if half the object-glass be covered, half the field of view will be cut off, and the remaining half will be unaffected.

The *opera-glass*, single or binocular, is a Galilean telescope, or a pair of Galilean telescopes. In the best instruments, both object-glass and eye-glass are achromatic combinations of three pieces, as shown in section in the figure (Fig. 749); the middle piece in each case being flint, and the other two crown (§ 1064).

1034. Reflecting Telescopes.—In reflecting telescopes, the place of an object-glass is supplied by a concave mirror called a *speculum*, usually composed of solid metal. The real and inverted image which it forms of distant objects is, in the Herschelian telescope, viewed directly through an eye-piece, the back of the observer being towards the object, and his face towards the speculum. This construction is only applicable to very large specula; as in instruments of ordinary

size the interposition of the observer's head would occasion too serious a loss of light.

An arrangement more frequently adopted is that devised by Sir Isaac Newton, and employed by him in the first reflecting telescope ever constructed. It is represented in Fig. 750. The speculum is at the bottom of a tube whose open end is directed towards the distant object to be examined. The rays 1 and 2 from one extremity of the object are reflected towards a, and the rays 3, 4 from the other extremity are reflected towards b. A real inverted image $a\,b$ would thus be formed at the principal focus of the concave speculum; but a small plane mirror M is interposed obliquely, and causes the real image to be formed at $a'\,b'$ in a symmetrical position with respect to the mirror M. The eye-lens O′ transforms this into the enlarged and virtual image A′B′.

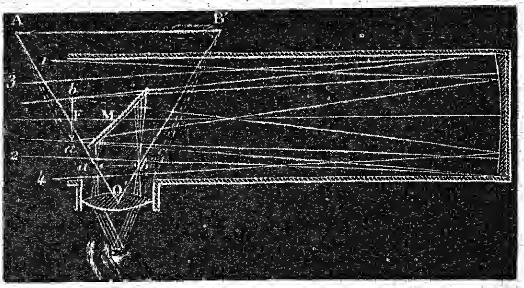


Fig. 750.—Newtonian Telescope.

Magnifying Power.—The approximate formula for the magnifying power is the same as in the case of the refracting telescopes already described. In fact the first image $a\,b$ subtends, at the optical centre O (not shown in the figure) of the large speculum, an angle $a\,O\,b$ equal to the visual angle for the naked eye; and the second image $a'\,b'$ (which is equal to the former) subtends, at the centre of the eye-piece, an angle $a'\,O'\,b'$ equal to the angle under which the image is seen in the telescope. The magnifying power is therefore $\frac{a'\,O'\,b'}{a\,O\,b}$, or, what is the same thing, is the ratio of the distance of $a\,b$ from O to the distance of $a'\,b'$ from O′, or the ratio of the focal length of the speculum to that of the eye-piece.

In the Gregorian telescope, which was invented before that of Newton, but not manufactured till a later date, there are two concave specula. The large one, which receives the direct rays from the object, forms a real and inverted image. The smaller speculum which is suspended in the centre of the tube, with its back to the object, receives the rays reflected from the first speculum, and forms a second real image, which is the enlarged and inverted image of the

first, and is therefore erect as compared with the object. This real and erect image is then magnified by means of an eye-piece, as in the instruments previously described, the eye-piece being contained in a tube which slides in a hole pierced in the middle of the large speculum.

As this arrangement gives an erect image, and enables the observer to look directly towards the object, it is specially convenient for terrestrial observation. It is the construction almost universally adopted in reflecting telescopes of small size.

The Cassegranian telescope resembles the Gregorian, except that the second speculum is convex, and the image which the observer sees is inverted.

1035. Silvered Specula.—Achromatic refracting telescopes give much better results, both as regards light and definition, than reflectors of the same size or weight; but it has been found practicable to make specula of much larger size than object-glasses. The aperture of Lord Rosse's largest telescope is 6 feet, whereas the aperture of the largest achromatic telescopes yet constructed is less than two feet, and increase of size involves increased thickness of glass, and consequent absorption of light.

The massiveness which is found necessary in the speculum in order to prevent flexure, is a serious inconvenience, as is also the necessity for frequent repolishing—an operation of great delicacy, as the slightest change in the form of the surface impairs the definition of the images. Both these defects have been to a certain extent remedied by the introduction of glass specula, covered in front with a thin coating of silver. Glass is much more easily worked than speculum-metal (which is remarkable for its brittleness in casting), and has only one-third of its specific gravity. Silver is also much superior to speculum-metal in reflecting power; and as often as it becomes tarnished it can be removed and renewed, without liability to change of form in the reflecting surface.[1]

1036. Measure of Brightness.—The apparent brightness of a surface is most naturally measured by the amount of light per unit area of its image on the retina; and therefore varies *directly as the amount of light which the surface sends to the pupil, and inversely as the apparent area of the surface.*

To avoid complications arising from the varying condition of the

[1] The merits of silvered specula are fully set forth in a brochure published by Mr. Browning, the optician, entitled *A Plea for Reflectors.*

observer, we shall leave dilatation and contraction of the pupil out of account.

When a body is looked at through a small pinhole in a card held close to the eye, it appears much darker than when viewed in the ordinary way; and in like manner images formed by optical instruments often furnish beams of light too narrow to fill the pupil. In all such cases it becomes necessary to distinguish between *effective brightness* and *intrinsic brightness*, the former being less than the latter in the same ratio in which the cross section of the beam which enters the pupil is less than the area of the pupil. We may correctly speak of the *intrinsic brightness* of a surface *for a particular point* of the pupil; and the effective brightness will in every case be the average value of the intrinsic brightness taken over the whole pupil.

In the case of natural bodies viewed in the ordinary way, the distinction may be neglected, since they usually send light in sensibly equal amounts to all parts of the pupil.

To obtain a numerical measure of intrinsic brightness, let us denote by s a small area on a surface directly facing towards the eye, or the foreshortened projection of a small area oblique to the line of vision, and by ω the solid angle which the pupil of the eye subtends at any point of s. Then the quantity of light q which s sends to the pupil per unit time, varies jointly as the solid angle ω, the area s, and the intrinsic brightness of s, which we will denote by I. We may therefore write

$$q = \mathrm{I}\, s\, \omega, \text{ and } \mathrm{I} = \frac{q}{s\,\omega}.$$

The intrinsic brightness of a small area s is therefore measured by $\frac{q}{s\,\omega}$, where q denotes the quantity of light which s emits per unit time in directions limited by the small angle of divergence ω.

1037. Applications.—One of the most obvious consequences is that *surfaces appear equally bright at all distances* in the same direction, provided that no light is stopped by the air or other intervening medium; for q and ω both vary inversely as the square of the distance. The area of the image formed on the retina in fact varies directly as the amount of light by which it is formed.

Images formed by Lenses.—Let A B (Fig. 751) be an object, and $a\,b$ its real image formed by the lens C D, whose centre is O. Let S denote a small area at A, and Q the quantity of light which it sends to the lens; also let s denote the corresponding area of the

image, and q the quantity of light which traverses it. Then q would be identical with Q if no light were stopped by the lens; the areas, S, s, are directly as the squares of the conjugate focal distances O A, O a; and the solid angles of divergence Ω and ω, for Q and q, being the solid angles subtended by the lens at A and a (for the plane angle $c\,a\,d$ in the figure is equal to the vertical angle $\mathrm{C}a\mathrm{D}$), are inversely as the squares of the conjugate focal distances. We have accordingly $\mathrm{S}\,\Omega = s\,\omega$. The intrinsic brightness $\frac{q}{s\,\omega}$ of the image is therefore equal to the intrinsic brightness $\frac{\mathrm{Q}}{\mathrm{S}\,\Omega}$ of the object except in so far as light is stopped by the lens. Precisely similar reasoning applies to virtual images formed by lenses.[1]

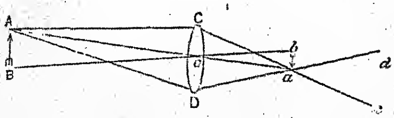


Fig. 751.—Brightness of Image.

In the case of images formed by mirrors, Ω and ω are the solid angles subtended by the mirror at the conjugate foci, and are inversely as the squares of the distances from the mirror; while S and s are directly as the squares of the distances from the centre of curvature; but these four distances are proportional (§ 967), so that the same reasoning is still applicable. If the mirror only reflects half the incident light, the image will have only half the intrinsic brightness of the object.

If the pupil is filled with light from the image, the effective brightness will be the same as the intrinsic brightness thus computed. If this condition is not fulfilled, the former will be less than the latter. When the image is greatly magnified as compared with the object, the angle of divergence is greatly diminished in comparison with the angle which the lens or mirror subtends at the object, and often becomes so small that only a small part of the pupil is utilized. This is the explanation of the great falling off of light which is observed in the use of high magnifying powers, both in microscopes and telescopes.

[1] For refraction out of a medium of index μ_1 into another of index μ_2, we have by § 1015, equation (13), $\mu_1 : \mu_2 :: \frac{\mathrm{A\,P_1}}{\mathrm{C\,P_1}} : \frac{\mathrm{A\,P_2}}{\mathrm{C\,P_2}}$. But since s_1, s_2 are the areas of corresponding parts of object and image, we have $s_1 : s_2 :: \mathrm{C\,P_1}^2 : \mathrm{C\,P_2}^2$, and since ω_1, ω_2 are the solid angles subtended at $\mathrm{P_1}$, $\mathrm{P_2}$ by one and the same portion of the bounding surface, we have $\omega_1 : \omega_2 :: \mathrm{A\,P_2}^2 : \mathrm{A\,P_1}^2$. Therefore $\frac{q}{s_1\,\omega_1} : \frac{q}{s_2\,\omega_2} :: \mu_1^2 : \mu_2^2$. The intrinsic brightnesses of a succession of images in different media are therefore directly as the squares of the absolute indices.

1038. Brightness of Image in a Telescope.—It has been already pointed out (§ 1031) that in most forms of telescope (the Galilean being an exception), there is a certain position, a little behind the eye-piece, at which a well-defined bright spot is formed upon a screen held there while the telescope is directed to any distant source of light. It has also been pointed out that this spot is the image, formed by the eye-piece, of the opening which is filled by the object-glass, and that the magnifying power of the instrument is the ratio of the size of the object-glass to the size of this bright spot.

Let s denote the diameter of the bright spot, o the diameter of the object-glass, e the diameter of the pupil of the eye; then $\frac{o}{s}$ is the linear magnifying power.

We shall first consider the case in which the spot exactly covers the pupil of the observer's eye, so that $s=e$. Then the whole light which traverses the telescope from a distant object enters the eye; and if we neglect the light stopped in the telescope, this is the whole light sent by the object to the object-glass, and is $\left(\frac{o}{e}\right)^2$ times that which would be received by the naked eye. The magnification of apparent area is $\left(\frac{o}{s}\right)^2$, which, from the equality of s and e, is the same as the increase of total light. The brightness is therefore the same as to the naked eye.

Next, let s be greater than e, and let the pupil occupy the central part of the spot. Then, since the spot is the image of the object-glass, we may divide the object-glass into two parts—a central part whose image coincides with the pupil, and a circumferential part whose image surrounds the pupil. All rays from the object which traverse the central part, traverse its image, and therefore enter the pupil; whereas rays traversing the circumferential part of the object-glass, traverse the circumferential part of the image, and so are wasted. The area of the central part (whether of the object-glass or of its image) is to the whole area as $e^2 : s^2$; and the light which the object sends to the central portion, instead of being $\left(\frac{o}{e}\right)^2$ times that which would be received by the naked eye, is only $\left(\frac{o}{s}\right)^2$ times. But $\left(\frac{o}{s}\right)^2$ is the magnification of apparent area. Hence the brightness is the same as to the naked eye. In these two cases, effective and intrinsic brightness are the same.

Lastly (and this is by far the most common case in practice), let s be less than e. Then no light is wasted, but the pupil is not filled. The light received is $\left(\frac{o}{e}\right)^2$ times that which the naked eye would receive; and the magnification of apparent area is $\left(\frac{o}{s}\right)^2$. The effective brightness of the image, is to the brightness of the object to the naked eye, as $\left(\frac{o}{e}\right)^2 : \left(\frac{o}{s}\right)^2$; that is, as $s^2 : e^2$; that is, as the area of the bright spot to the whole area of the pupil.

To correct for the light stopped by reflection and imperfect transparency, we have simply to multiply the result in each case by a proper fraction, expressing the ratio of the transmitted to the incident light. This ratio, for the central parts of the field of view, is about 0·85 in the best achromatic telescopes. In such telescopes, therefore, the brightness of the image cannot exceed 0·85 of the brightness of the object to the naked eye. It will have this precise value, when the magnifying power is equal to or less than $\frac{o}{e}$; and from this point upwards will vary inversely as the square of the linear magnification.

The same formulæ apply to reflecting telescopes, o denoting now the diameter of the large speculum which serves as objective; but the constant factor is usually considerably less than 0·85.

It may be accepted as a general principle in optics, that while it is possible, by bad focussing or instrumental imperfections, to obtain a confused image whose brightness shall be intermediate between the brightest and the darkest parts of the object, *it is impossible, by any optical arrangement whatever, to obtain an image whose brightest part shall surpass the brightest part of the object.*

1039. Brightness of Stars.—There is one important case in which the foregoing rules regarding the brightness of images become nugatory. The fixed stars are bodies which subtend at the earth angles smaller than the *minimum visibile*, but which, on account of their excessive brightness, *appear* to have a sensible angular diameter. This is an instance of *irradiation*, a phenomenon manifested by all bodies of excessive brightness, and consisting in an extension of their apparent beyond their actual boundary. What is called, in popular language, a bright star, is a star which sends a large total amount of light to the eye.

Denoting by a the ratio of the transmitted to the whole incident light, a ratio which, as we have seen, is about 0·85 in the most

favourable cases, and calling the light which a star sends to the naked eye unity, the light perceived in its image will be $\alpha\left(\frac{o}{s}\right)^2$, or $\alpha \times$ square of linear magnification, if the bright spot is as large as the pupil. When the eye-piece is changed, increase of power diminishes the size of the spot, and increases the light received by the eye, until the spot is reduced to the size of the pupil. After this, any further magnification has no effect on the quantity of light received, its constant value being $\alpha\left(\frac{o}{e}\right)^2$.

The value of this last expression, or rather the value of $\alpha\, o^2$, is the measure of what is called the *space-penetrating power* of a telescope; that is to say, the power of rendering very faint stars visible; and it is in this respect that telescopes of very large aperture, notably the great reflector of Lord Rosse, are able to display their great superiority over instruments of moderate dimensions.

We have seen that the total light in the visible image of a star remains unaltered, by increase of power in the eye-piece above a certain limit. But the visibility of faint stars in a telescope is promoted by darkening the back-ground of sky on which they are seen. Now the brightness of this back-ground varies directly as s^2, or inversely as the square of the linear magnification (s being supposed less than e). Hence it is advantageous, in examining very faint stars, to employ eye-pieces of sufficient power to render the bright spot much smaller than the pupil of the eye.

1040. Images on a Screen.—Thus far we have been speaking of the brightness of images as viewed directly. Images cast upon a screen are, as a matter of fact, much less brilliant; partly because the screen sends out light in all directions, and therefore through a much larger solid angle than that formed by the beam incident on the screen, and partly because some of the incident light is absorbed.

Let A be the area of the object, which we suppose to face directly towards the lens by which the image is thrown upon the screen, a the area of the image, and D, d their respective distances from the lens. Then if I denote the intrinsic brightness of the object, the light sent from A to the lens will be the product of I A by the solid angle which the lens subtends at the object. This solid angle will be $\frac{L}{D^2}$, if L denote the area of the lens. I A $\frac{L}{D^2}$ is therefore the light sent by the object to the lens, and if we neglect reflection and absorption all this light falls upon the image. *The light which falls*

on unit area of the image is therefore $I \frac{A}{a} \frac{L}{D^2}$, that is $I \frac{L}{d^2}$; it *is* therefore *the same as if the lens were a source of light of brightness I.* Accordingly, if the image of a lamp flame be thrown upon the pupil of an observer's eye, and be large enough to cover the pupil, he will see the lens filled with light of a brightness equal to that of the flame seen directly.

1041. Field of View in Astronomical Telescope.—Let $p\,m\,n\,q$ (Fig. 752) be the common focal plane of the object-glass and eye-glass. Draw $B\,a$, $A\,b$ joining the highest points of both, and the lowest points of both; also $A\,a$, $B\,b$ joining the highest point of each with the lowest point of the other. Evidently $B\,a$, $A\,b$ will be the boundaries of the beam of light transmitted through the telescope, and therefore the points p and q in which these lines intersect the focal plane, will be the extremities of that part of the real image which sends rays to the eye. The angle subtended by $p\,q$ at the centre of the object-glass will therefore be the angular diameter of the complete field of view. But the outer portions of this field will be less bright than the centre, and the full amount of brightness, as calculated in § 1038 for the case in which the "bright spot" is smaller than the pupil, will belong only to the portion $m\,n$ bounded by the cross-lines $A\,a$, $B\,b$; for all the rays sent by the object-glass through the part $m\,n$ traverse the eye-glass, and therefore the bright spot, whereas some of the rays sent by the object-glass to any point between m and p, or between n and q pass wide of the eye-glass and therefore do not reach the bright spot. The complete field of view, as seen by an eye whose pupil includes the bright spot, accordingly consists of a central disc $m\,n$ of full brightness, surrounded by a ring extending to p and q whose brightness gradually diminishes from full brightness at its junction with the disc to zero at its outer boundary. This ring is called the "ragged edge," and is put out of sight in actual telescopes by an opaque stop of annular form in the focal plane. The angular diameter of the field of view, excluding the ragged edge, will be equal to the angle which $m\,n$ subtends at the centre of the object-glass.

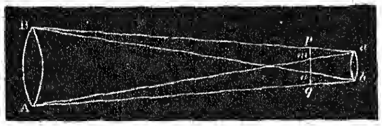


Fig. 752.—Field of View.

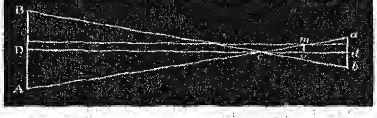


Fig. 753.—Calculation of Field.

To calculate the length of mn, join D, d, the centres of the object-glass and eye-glass (Fig. 753). The joining line will obviously pass through the intersection of Aa, Bb, and also through the middle point of mn. Draw a parallel to this line through m. Then, by comparing the similar triangles of which am, Am are the hypotenuses, we have

$$ad - mo : od :: \text{AD} + mo : \text{D}o.$$

Hence, multiplying extremes and means, and denoting the focal lengths Do, od by F, f, we have

$$\text{F}\,(ad - mo) = f\,(\text{AD} + mo),$$

whence

$$mo = \frac{\text{F}\,.\,ad - f\,.\,\text{AD}}{\text{F} + f}.$$

This is the radius of the real image, excluding the ragged edge; and the angular radius of the field of view will be

$$\frac{mo}{\text{F}} = \frac{\text{F}\,.\,ad - f\,.\,\text{AD}}{\text{F}\,(\text{F}+f)}$$
$$= \frac{ad}{\text{F}+f} - \frac{f\,.\,\text{AD}}{\text{F}\,(\text{F}+f)}.$$

The first term $\frac{ad}{\text{F}+f}$ is the angle which the radius of the eye-glass subtends at the object-glass. But, it is obvious from Fig. 752 that the line aD would bisect mp. Hence the second term represents half the breadth of the ragged edge, and the whole field of view, including the ragged edge, has an angular radius

$$\frac{ad}{\text{F}+f} + \frac{f\,.\,\text{AD}}{\text{F}\,(\text{F}+f)}.$$

1042. Cross-wires of Telescopes.—We have described in § 1010 a mode of marking the place of a real image by means of a cross of threads. When telescopes are employed to assist in the measurement of angles, a contrivance of this kind is almost always introduced. A cross of silkworm threads, in instruments of low power, or of spider threads in instruments of higher power, is stretched across a metallic frame just in front of the eye-piece. The observer must first adjust the eye-piece for distinct vision of this cross, and must then (in the case of theodolites and other surveying instruments) adjust the distance of the object-glass until the object which is to be observed is also seen distinctly in the telescope. The image of the object will then be very nearly in the plane of the cross. If it is not exactly in the plane, parallactic displacement will be observed when the eye is shifted, and this must be cured by slightly

altering the distance of the object-glass. When the adjustment has been completed, the cross always marks one definite point of the object, however the eye be shifted. This coincidence will not be disturbed by pushing in or pulling out the eye-piece; for the frame which carries the cross is attached to the body of the telescope, and the coincidence of the cross with a point of the image is real, so that it could be observed by the naked eye, if the eye-piece were removed. The adjustment of the eye-piece merely serves to give distinct vision, and this will be obtained simultaneously for both the cross and the object.

1043. Line of Collimation.—The employment of *cross-wires* (as these crossing threads are called) enormously increases our power of making accurate observations of direction, and constitutes one of the greatest advantages of modern over ancient instruments.

The line which is regarded as the line of sight, or as the direction in which the telescope is pointed, is called the *line of collimation.* If we neglect the curvature of rays due to atmospheric refraction, we may define it as the *line joining the cross to the object whose image falls on it.* More rigorously, the line of collimation is the *line joining the cross to the optical centre of the object-glass.* When it is desired to adjust the line of collimation,—for example, to make it truly perpendicular to the horizontal axis on which the telescope is mounted, the adjustment is performed by shifting the frame which carries the wires, slow-motion screws being provided for this purpose. Telescopes for astronomical observation are often furnished with a number of parallel wires, crossed by one or two in the transverse direction; and the line of collimation is then defined by reference to an imaginary cross, which is the centre of mean position of all the actual crosses.

1044. Micrometers.—Astronomical micrometers are of various kinds, some of them serving for measuring the angular distance between two points in the same field of view, and others for measuring their apparent direction from one another. They often consist of spider threads placed in the principal focus of the object-glass, so as to be in the same plane as the images of celestial objects, one or more of the threads being movable by means of slow-motion screws, furnished with graduated circles, on which parts of a turn can be read off.

One of the commonest kinds consists of two parallel threads, which can thus be moved to any distance apart, and can also be turned round in their own plane.

CHAPTER LXXII.

DISPERSION. STUDY OF SPECTRA.

1045. Newtonian Experiment.—In the chapter on refraction, we have postponed the discussion of one important phenomenon by which it is usually accompanied, and which we must now proceed to explain. The following experiment, which is due to Sir Isaac Newton, will furnish a fitting introduction to the subject.

On an extensive background of black, let three bright strips be laid in line, as in the left-hand part of Fig. 754, and looked at through a prism with its refracting edge parallel to the strips. We

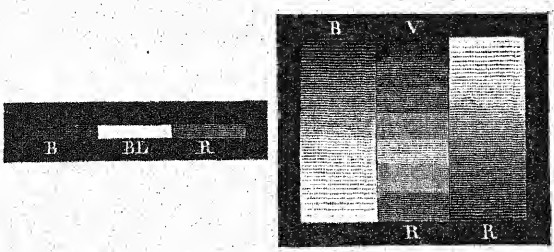


Fig. 754.—Spectra of White and Coloured Strips.

shall suppose the edge to be upward, so that the image is raised above the object. The images, as represented in the right-hand part of Fig. 754, will have the same horizontal dimensions as the strips, but will be greatly extended in the vertical direction; and each image, instead of having the uniform colour of the strips from which it is derived, will be tinted with a gradual succession of colours from top to bottom. Such images are called *spectra*.

If one of the strips (the middle one in the figure) be white, its spectrum will contain the following series of colours, beginning at the top: *violet, blue, green, yellow, orange, red.*

If one of the strips be blue (the left-hand one in the figure), its image will present bright colours at the upper end; and these will be identical with the colours adjacent to them in the spectrum of white. The colours which form the lower part of the spectrum of white will either be very dim and dark in the spectrum of blue, or will be wanting altogether, being replaced by black.

If the other strip be red, its image will contain bright colours at the lower or red end, and those which belong to the upper end of the spectrum of white will be dim or absent. Every colour that occurs in the spectrum of blue or of red will also be found, and in the same horizontal line, in the spectrum of white.

If we employ other colours instead of blue or red, we shall obtain analogous results; every colour will be found to give a spectrum which is identical with part of the spectrum of white, both as regards colour and position, but not generally as regards brightness.

We may occasionally meet with a body whose spectrum consists only of one colour. The petals of some kinds of convolvulus give a spectrum consisting only of blue, and the petals of nasturtium give only red.

1046. Composite Nature of Ordinary Colours.—This experiment shows that the colours presented by the great majority of natural bodies are composite. When a colour is looked at with the naked eye, the sensation experienced is the joint effect of the various elementary colours which compose it. The prism serves to resolve the colour into its components, and exhibit them separately. The experiment also shows that a mixture of all the elementary colours in proper proportions produces white.

1047. Solar Spectrum.—The coloured strips in the foregoing experiment may be illuminated either by daylight or by any of the ordinary sources of artificial light. The former is the best, as gas-light and candle-light are very deficient in blue and violet rays.

Colour, regarded as a property of a coloured (opaque) body, is the power of selecting certain rays and reflecting them either exclusively or in larger proportion than others. The spectrum presented by a body viewed by reflected light, as ordinary bodies are, can thus only consist of the rays, or a selection of the rays, by which the body is illuminated.

A beam of solar light can be directly resolved into its constituents by the following experiment, which is also due to Newton, and was the first demonstration of the composite character of solar light.

Let a beam of sun-light be admitted through a small opening into a dark room. If allowed to fall normally on a white screen, it produces (§ 938) a round white spot, which is an image of the sun. Now let a prism be placed in its path edge-downwards, as in Fig. 755; the

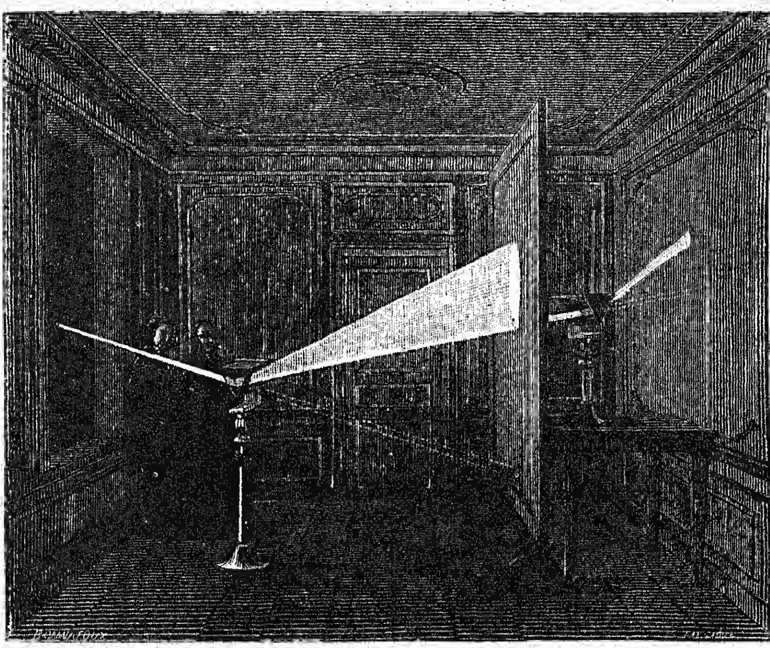


Fig. 755.—Solar Spectrum by Newton's Method.

beam will thus be deflected upwards, and at the same time resolved into its component colours. The image depicted on the screen will be a many-coloured band, resembling the spectrum of white described in § 1045. It will be of uniform width, and rounded off at the ends, being in fact built up of a number of overlapping discs, one for each kind of elementary ray. It is called the *solar spectrum.*

The rays which have undergone the greatest deviation are the violet. They occupy the upper end of the spectrum in the figure. Those which have undergone the least deviation are the red. Of all visible rays, the violet are the most, and the red the least refrangible; and the analysis of light into its components by means of the prism is due to difference of refrangibility. If a small opening is made in the screen, so as to allow rays of only one colour to pass, it will be found,

on transmitting these through a second prism behind the screen, as in Fig. 755, that no further analysis can be effected, and the whole of the image formed by receiving this transmitted light on a second screen will be of this one colour.

1048. Mode of obtaining a Pure Spectrum.—The spectra obtained by the methods above described are built up of a number of overlapping images of different colours. To prevent this overlapping, and obtain each elementary colour pure from all admixture with the rest, we must in the first place employ as the object for yielding the images a very narrow line; and in the second place we must take care that the images which we obtain of this line are not blurred, but have the greatest possible sharpness. A spectrum possessing these characteristics is called pure.

The simplest mode of obtaining a pure spectrum consists in looking through a prism at a fine slit in the shutter of a dark room. The edges of the prism must be parallel to the slit, and its distance from the slit should be five feet or upwards. The observer, placing his eye close to the prism, will see a spectrum; and he should rotate the prism on its axis until he has brought this spectrum to its smallest angular distance from the real slit, of which it is the image.

Let E (Fig. 756) be the position of the eye, S that of the slit. Then the extreme red and violet images of the slit will be seen at R, V, at distances from the prism sensibly equal to the real distance of S (§ 997); and the other images, which compose the remainder of the spectrum, will occupy positions between R and V. The spectrum, in this mode of operating, is virtual.

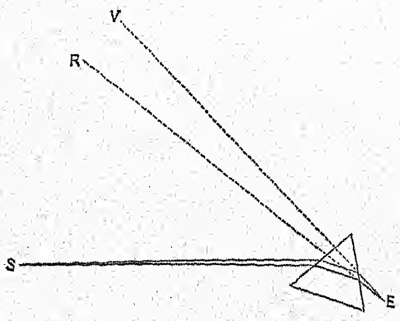


Fig. 756.—Arrangement for seeing a Pure Spectrum.

To obtain a real spectrum in a state of purity, a convex lens must be employed. Let the lens L (Fig. 757) be first placed in such a position as to throw a sharp image of the slit S upon a screen at I. Next let a prism P be introduced between the lens and screen, and rotated on its axis till the position of minimum deviation is obtained, as shown by the movements of the impure spectrum which travels about the walls of the room. Then if the screen be moved into the position R V, its distance from the prism being the

same as before, a pure spectrum will be depicted upon it. A similar result can be obtained by placing the prism between the lens and the slit, but the adjustments are rather more troublesome. Direct

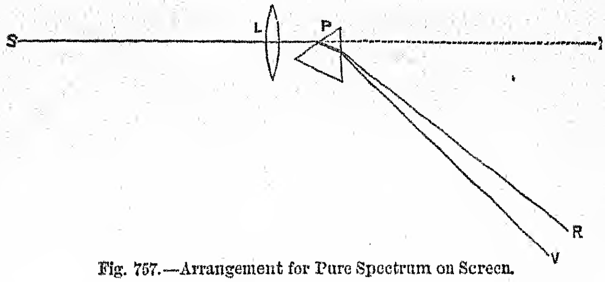


Fig. 757.—Arrangement for Pure Spectrum on Screen.

sun-light, or sun-light reflected from a mirror placed outside the shutter, is necessary for this experiment, as sky-light is not sufficiently powerful. It is usual, in experiments of this kind, to employ a movable mirror called a *heliostat*, by means of which the light can be reflected in any required direction. Sometimes the movements of the mirror are obtained by hand; sometimes by an ingenious clock-work arrangement, which causes the reflected beam to keep its direction unchanged notwithstanding the progress of the sun through the heavens.

The advantage of placing the prism in the position of minimum deviation is twofold. First, the adjustments are facilitated by the equality of conjugate focal distances, which subsists in this case and in this only. Secondly and chiefly, this is the only position in which the images are not blurred. In any other position it can be shown[1] that a small cone of homogeneous incident rays is no longer a cone (that is, its rays do not accurately pass through one point) after transmission through the prism.

The method of observation just described was employed by Wollaston, in the earliest observations of a pure spectrum ever obtained. Fraunhofer, a few years later, independently devised the same method, and carried it to much greater perfection. Instead of looking at the virtual image with the naked eye, he viewed it through a telescope, which greatly magnified it, and revealed several features never before detected. The prism and telescope were at a distance of 24 feet from the slit.

[1] Parkinson's *Optics*, § 96. Cor. 2.

47 **1049. Dark Lines in the Solar Spectrum.**—When a pure spectrum of solar light is examined by any of these methods, it is seen to be traversed by numerous dark lines, constituting, if we may so say, dark images of the slit. Each of these is an indication that a particular kind of elementary ray is wanting[1] in solar light. Every elementary ray that is present gives its own image of the slit in its own peculiar colour; and these images are arranged in strict contiguity, so as to form a continuous band of light passing by perfectly gradual transitions through the whole range of simple colour, except at the narrow intervals occupied by the dark lines. Fig. 1, Plate III., is a rough representation of the appearance thus presented. If the slit is illuminated by a gas flame, or by any ordinary lamp, instead of by solar light, no such lines are seen, but a perfectly continuous spectrum is obtained. The dark lines are therefore not characteristic of light in general, but only of solar light.

Wollaston saw and described some of the more conspicuous of them. Fraunhofer counted about 600, and marked the places of 354 upon a map of the spectrum, distinguishing some of the more conspicuous by the names of letters of the alphabet, as indicated in fig. 1. These lines are constantly referred to as reference marks for the accurate specification of different portions of the spectrum. They always occur in precisely the same places as regards colour, but do not retain exactly the same relative distances one from another, when prisms of different materials are employed, different parts of the spectrum being unequally expanded by different refracting substances.[2] The inequality, however, is not so great as to introduce any difficulty in the identification of the lines.

The dark lines in the solar spectrum are often called Fraunhofer's lines. Fraunhofer himself called them the "fixed lines."

47 **1050. Invisible Rays of the Spectrum.**—The brightness of the solar spectrum, however obtained, is by no means equal throughout, but is greatest between the dark lines D and E; that is to say, in the yellow and the neighbouring colours orange and light green; and falls off gradually on both sides.

The heating effect upon a small thermometer or thermopile increases in going from the violet to the red, and still continues to increase for a certain distance beyond the visible spectrum at the red end. Prisms and lenses of rock-salt should be employed for this

[1] Probably not absolutely wanting, but so feeble as to appear black by contrast.

[2] This property is called the *irrationality of dispersion*.

investigation, as glass largely absorbs the invisible rays which lie beyond the red.

When the spectrum is thrown upon the sensitized paper employed in photography, the action is very feeble in the red, strong in the blue and violet, and is sensible to a great distance beyond the violet end. When proper precautions are taken to insure a very pure spectrum, the photograph reveals the existence of dark lines, like those of Fraunhofer, in the invisible ultra-violet portion of the spectrum. The strongest of these have been named L, M, N, O, P. Quiz

1051. Phosphorescence and Fluorescence.— There are some substances which, after being exposed in the sun, are found for a long time to appear self-luminous when viewed in the dark, and this

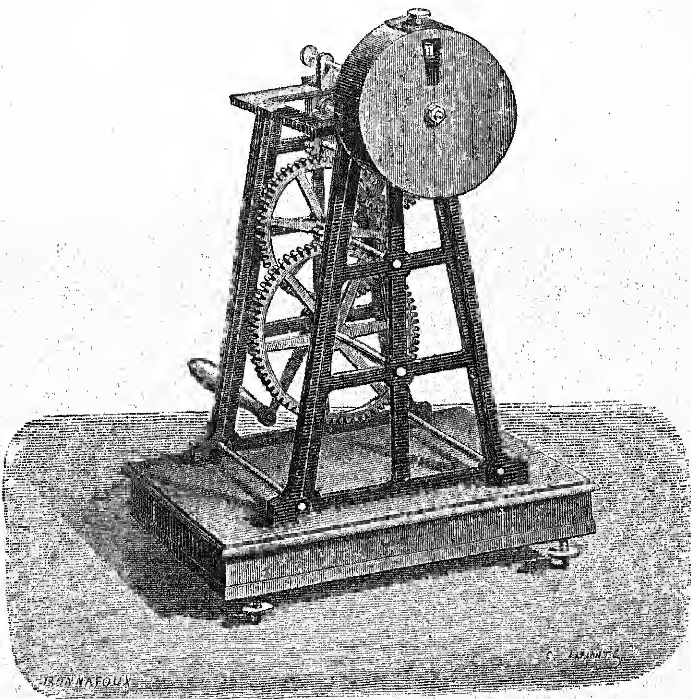


Fig. 758.—Becquerel's Phosphoroscope.

without any signs of combustion or sensible elevation of temperature. Such substances are called *phosphorescent.* Sulphuret of calcium and sulphuret of barium have long been noted for this property, and have hence been called respectively *Canton's phosphorus,* and *Bologna*

phosphorus. The phenomenon is chiefly due to the action of the violet and ultra-violet portion of the sun's rays.

More recent investigations have shown that the same property exists in a much lower degree in an immense number of bodies, their phosphorescence continuing, in most cases, only for a fraction of a second after their withdrawal from the sun's rays. E. Becquerel has contrived an instrument, called the *phosphoroscope*, which is extremely appropriate for the observation of this phenomenon. It is represented in Fig. 758. Its most characteristic feature is a pair of rigidly connected discs (Fig. 759), each pierced with four openings, those of the one being not opposite but midway between those of the other. This pair of discs can be set in very rapid rotation by means of a series of wheels and pinions. The body to be examined is attached to a fixed stand between the two discs, so that it is alternately exposed on opposite sides as the discs rotate. One side is turned towards the sun, and the other towards the observer, who accordingly only sees the body when it is not exposed to the sun's rays. The cylindrical case within which the discs revolve, is fitted into a hole in the shutter of a dark room, and is pierced with an opening on each side exactly opposite the position in which the body is fixed. The body, if not phosphorescent, will never be seen by the observer, as it is always in darkness except when it is hidden by the intervening disc. If its phosphorescence lasts as long as an eighth part of the time of one rotation, it will become visible in the darkness.

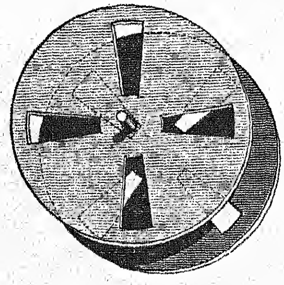


Fig. 759.
Discs of Phosphoroscope.

Nearly all bodies, when thus examined, show traces of phosphorescence, lasting, however, in some cases, only for a ten-thousandth of a second.

The phenomenon of *fluorescence*, which is illustrated in Plate II. accompanying § 817, appears to be essentially identical with phosphorescence. The former name is applied to the phenomenon, if it is observed while the body is actually exposed to the source of light, the latter to the effect of the same kind, but usually less intense, which is observed after the light from the source is cut off. Both forms of the phenomenon occur in a strongly-marked degree in the same bodies. Canary-glass, which is coloured with oxide of uranium, is

a very convenient material for the exhibition of fluorescence. A thick piece of it, held in the violet or ultra-violet portion of the solar spectrum, is filled to the depth of from $\frac{1}{8}$ to $\frac{1}{4}$ of an inch with a faint nebulous light. A solution of sulphate of quinine is also frequently employed for exhibiting the same effect, the luminosity in this case being bluish. If the solar spectrum be thrown upon a screen freshly washed with sulphate of quinine, the ultra-violet portion will become visible by fluorescence; and if the spectrum be very pure, the presence of dark lines in this portion will be detected.

The light of the electric lamp is particularly rich in ultra-violet rays, this portion of its spectrum being much longer than in the case of solar light, and about twice as long as the spectrum of luminous rays. Prisms and lenses of quartz should be employed for this purpose, as this material is specially transparent to the highly-refrangible rays. Flint-glass prisms, however, if of good quality, answer well in operating on solar light. The luminosity produced by fluorescence has sensibly the same tint in all parts of the spectrum in which it occurs, and depends upon the fluorescent substance employed. Prismatic analysis is not necessary to the exhibition of fluorescence. The phenomenon is very conspicuous when the electric discharge of a Holtz's machine or a Ruhmkorff's coil is passed near fluorescent substances, and it is faintly visible when these substances are examined in bright sunshine. The light emitted by a fluorescent substance is found by analysis not to be homogeneous, but to consist of rays having a wide range of refrangibility.

The ultra-violet rays, though usually styled invisible, are not altogether deserving of this title. By keeping all the rest of the spectrum out of sight, and carefully excluding all extraneous light, the eye is enabled to perceive these highly refrangible rays. Their colour is described as lavender-gray or bluish white, and has been attributed, with much appearance of probability, to fluorescence of the retina. The ultra-red rays, on the other hand, are never seen; but this may be owing to the fact, which has been established by experiment, that they are largely, if not entirely, absorbed before they can reach the retina.

1052. Recomposition of White Light.—The composite nature of white light can be established by actual synthesis. This can be done in several ways.

1. If a second prism, precisely similar to the first, but with its refracting edge turned the contrary way, is interposed in the path of

the coloured beam, very near its place of emergence from the first prism, the deviation produced by the second prism will be equal and opposite to that produced by the first, the two prisms will produce the effect of a parallel plate, and the image on the screen will be a white spot, nearly in the same position as if the prisms were removed.

2. Let a convex lens (Fig. 760) be interposed in the path of the coloured beam, in such a manner that it receives all the rays, and

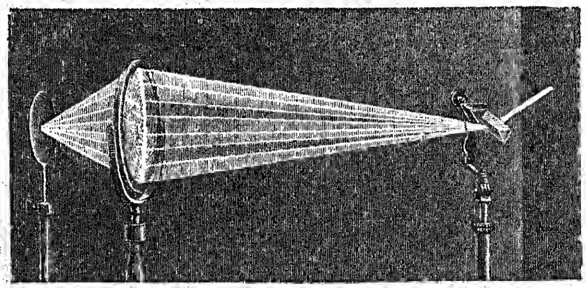


Fig. 760.—Recomposition by Lens.

that the screen and the prism are at conjugate focal distances. The image thus obtained on the screen will be white, at least in its central portions.

3. Let a number of plane mirrors be placed so as to receive the successive coloured rays, and to reflect them all to one point of a

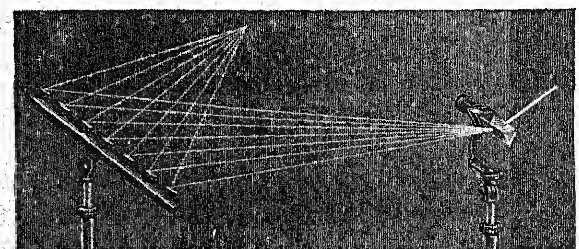


Fig. 761.—Recomposition by Mirrors.

screen, as in Fig. 761. The bright spot thus formed will be white or approximately white.

More complete information respecting the mixture of colours will be given in the next chapter.

√ 1053. **Spectroscope.**—When we have obtained a pure spectrum by any of the methods above indicated, we have in fact effected an analysis of the light with which the slit is illuminated. In recent years, many forms of apparatus have been constructed for this purpose, under the name of *spectroscopes*.

A spectroscope usually contains, besides a slit, a prism, and a telescope (as in Fraunhofer's method of observation), a convex lens called a *collimator*, which is fixed between the prism and the slit, at the distance of its principal focal length from the latter. The effect of this arrangement is, that rays from any point of the slit emerge parallel, as if they came from a much larger slit (the virtual image of the real slit) at a much greater distance. The prism (set at minimum deviation) forms a virtual image of this image at the same distance, but in a different direction, on the principle of Fig. 757.

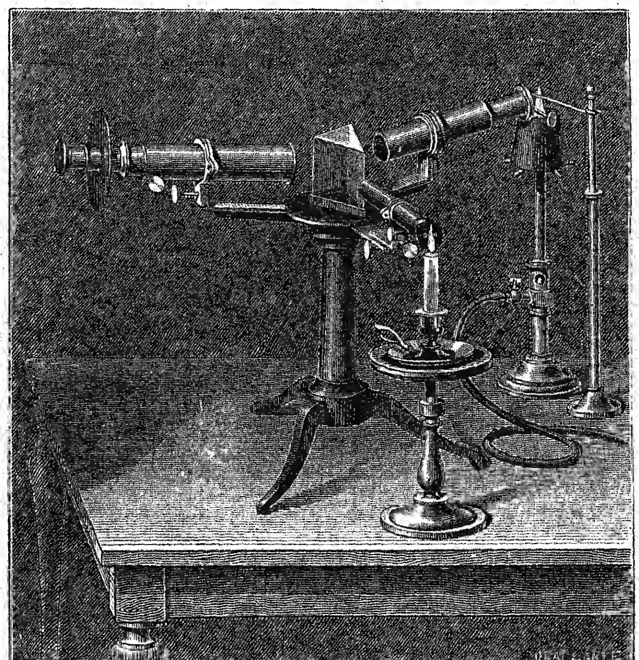


Fig. 762.—Spectroscope.

To this second virtual image the telescope is directed, being focussed as if for a very distant object.

Fig. 762 represents a spectroscope thus constructed. The tube of

the collimator is the further tube in the figure, the lens being at the end of the tube next the prism, while at the far end, close to the lamp flame, there is a slit (not visible in the figure) consisting of an opening between two parallel knife-edges, one of which can be moved to or from the other by turning a screw. The knife-edges must be very true, both as regards straightness and parallelism, as it is often necessary to make the slit exceedingly narrow. The tube on the left hand is the telescope, furnished with a broad guard to screen the eye from extraneous light. The near tube, with a candle opposite its end, is for purposes of measurement. It contains, at the end next the candle, a scale of equal parts, engraved or photographed on glass. At the other end of the tube is a collimating lens, at the distance of its own focal length from the scale; and the collimator is set so that its axis and the axis of the telescope make equal angles with the near face of the prism. The observer thus sees in the telescope, by reflection from the surface of the prism, a magnified image of the scale, serving as a standard of reference for assigning the positions of the lines in any spectrum which may be under examination. This arrangement affords great facilities for rapid observation.

Another plan is, for the arm which carries the telescope to be movable round a graduated circle, the telescope being furnished with cross-wires, which the observer must bring into coincidence with any line whose position he desires to measure.

Arrangements are frequently made for seeing the spectra of two different sources of light in the same field of view, one half of the length of the slit being illuminated by the direct rays of one of the sources, while a reflector, placed opposite the other half of the slit, supplies it with reflected light derived from the other source. This method should always be employed when there is a question as to the exact coincidence of lines in the two spectra. The reflector is usually an equilateral prism. The light enters normally at one of its faces, is totally reflected at another, and emerges normally at the third, as in the annexed sketch (Fig. 763), where the dotted line represents the path of a ray.

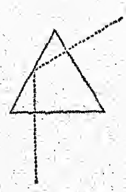


Fig. 763. Reflecting Prism.

A one-prism spectroscope is amply sufficient for the ordinary purposes of chemistry. For some astronomical applications a much greater dispersion is required. This is attained by making the light pass through a number of prisms in succession, each being set in the proper position for giving minimum deviation to the rays which have

passed through its predecessor. Fig. 764 represents the ground plan of such a battery of prisms, and shows the gradually increasing width of a pencil as it passes round the series of nine prisms on its way from the collimator to the telescope. The prisms are usually connected by a special arrangement, which enables the observer, by a single movement, to bring all the prisms at once into the proper position for giving minimum deviation to the particular ray under examination, a position which differs considerably for rays of different refrangibilities.

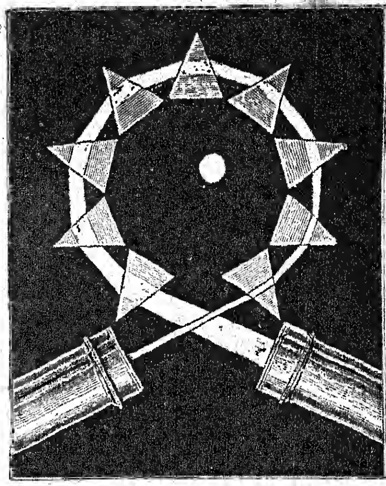


Fig. 764.—Train of Prisms.

1054. Use of Collimator.—The introduction of a collimating lens, to be used in conjunction with a prism and observing telescope, is due to Professor Swan.[1] Fraunhofer employed no collimator; but his prism was at a distance of 24 feet from the slit, whereas a distance of less than 1 foot suffices when a collimator is used.

It is obvious that homogeneous light, coming from a point at the distance of a foot, and falling upon the whole of one face of a prism —say an inch in width, cannot all have the incidence proper for minimum deviation. Those rays which very nearly fulfil this condition, will concur in forming a tolerably sharp image, in the position which we have already indicated. The emergent rays taken as a whole, do not diverge from any one point, but are tangents to a virtual caustic (§ 974). An eye receiving any portion of these rays, will see an image in the direction of a tangent from the eye to the caustic; and this image will be the more blurred as the deviation is further from the minimum. When the naked eye is employed, and the prism is so adjusted that the centre of the pupil receives rays of minimum deviation, a distance of five or six feet between the prism and slit is sufficient to give a sharp image; but if we employ an observing telescope whose object-glass is five times larger in diameter than the pupil of the eye, we must increase the distance between the

[1] *Trans. Roy. Soc. Edinburgh*, 1847 and 1856.

prism and slit fivefold to obtain equally good definition. A collimating lens, if achromatic and of good quality, gives the advantage of good definition without inconvenient length.

When exact measures of deviation are required, it confers the further advantage of altogether dispensing with a very troublesome correction for parallax.

1055. Different Kinds of Spectra.—The examination of a great variety of sources of light has shown that spectra may be divided into the following classes:—

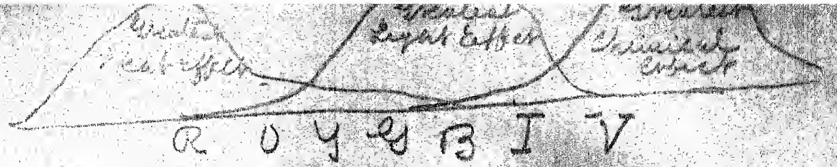
1. The solar spectrum is characterized, as already observed, by a definite system of dark lines interrupting an otherwise continuous succession of colours. The same system of dark lines is found in the spectra of the moon and planets, this being merely a consequence of the fact that they shine by the reflected light of the sun. The spectra of the fixed stars also contain systems of dark lines, which are different for different stars.

2. The spectra of incandescent solids and liquids are completely continuous, containing light of all refrangibilities from the extreme red to a higher limit depending on the temperature.

3. Flames not containing solid particles in suspension, but merely emitting the light of incandescent gases, give a discontinuous spectrum, consisting of a finite number of bright lines. The continuity of the spectrum of a gas or candle flame, arises from the fact that nearly all the light of the flame is emitted by incandescent particles of solid carbon,—particles which we can easily collect in the form of soot. When a gas-flame is fed with an excessive quantity of air, as in Bunsen's burner, the separation of the solid particles of carbon from the hydrogen with which they were combined, no longer takes place; the combustion is purely gaseous, and the spectrum of the flame is found to consist of bright lines. When the electric light is produced between metallic terminals, its spectrum contains bright lines due to the incandescent vapour of these metals, together with other bright lines due to the incandescence of the oxygen and nitrogen of the air. When it is taken between charcoal terminals, its spectrum is continuous; but if metallic particles be present, the bright lines due to their vapours can be seen as well.

The spectrum of the electric discharge in a Geissler's tube consists of bright lines characteristic of the gas contained in the tube.

1056. Spectrum Analysis.—As the spectrum exhibited by a compound substance when subjected to the action of heat, is frequently


Light Effect
R O Y G B I V

found to be identical with the spectrum of one of its constituents, or to consist of the spectra of its constituents superimposed,[1] the spectroscope affords an exceedingly ready method of performing qualitative analysis.

If a salt of a metal which is easily volatilized is introduced into a Bunsen lamp-flame, by means of a loop of platinum wire, the bright lines which form the spectrum of the metal will at once be seen in a spectroscope directed to the flame; and the spectrum of the Bunsen flame itself is too faint to introduce any confusion. For those metals which require a higher temperature to volatilize them, electric discharge is usually employed. Geissler's tubes are commonly used for gases.

Plate III. contains representations of the spectra of several of the more easily volatilized metals, as well as of phosphorus and hydrogen; and the solar spectrum is given at the top for comparison. The bright lines of some of these substances are precisely coincident with some of the dark lines in the solar spectrum.

The fact that certain substances when incandescent give definite bright lines, has been known for many years, from the researches of Brewster, Herschel, Talbot, and others; but it was for a long time thought that the same line might be produced by different substances, more especially as the bright yellow line of sodium was often seen in flames in which that metal was not supposed to be present. Professor Swan, having ascertained that the presence of the 2,500,000th part of a grain of sodium in a flame was sufficient to produce it, considered himself justified in asserting, in 1856, that this line was always to be taken as an indication of the presence of sodium in larger or smaller quantity.

But the greatest advance in spectral analysis was made by Bunsen and Kirchhoff, who, by means of a four-prism spectroscope, obtained accurate observations of the positions of the bright lines in the spectra of a great number of substances, as well as of the dark lines in the solar spectrum, and called attention to the identity of several of the latter with several of the former. Since the publication of their researches, the spectroscope has come into general use among chemists, and has already led to the discovery of four new metals, cæsium, rubidium, thallium, and indium.

1057. Reversal of Bright Lines. Analysis of the Sun's Atmosphere.

[1] These appear to be merely examples of the dissociation of the elements of a chemical compound at high temperatures.

found to be identical with the spectrum of one of its constituents, or to consist of the spectra of its constituents superimposed,[1] the spectroscope affords an exceedingly ready method of performing qualitative analysis.

If a salt of a metal which is easily volatilized is introduced into a Bunsen lamp-flame, by means of a loop of platinum wire, the bright lines which form the spectrum of the metal will at once be seen in a spectroscope directed to the flame; and the spectrum of the Bunsen flame itself is too faint to introduce any confusion. For those metals which require a higher temperature to volatilize them, electric discharge is usually employed. Geissler's tubes are commonly used for gases.

Plate III. contains representations of the spectra of several of the more easily volatilized metals, as well as of phosphorus and hydrogen; and the solar spectrum is given at the top for comparison. The bright lines of some of these substances are precisely coincident with some of the dark lines in the solar spectrum.

The fact that certain substances when incandescent give definite bright lines, has been known for many years, from the researches of Brewster, Herschel, Talbot, and others; but it was for a long time thought that the same line might be produced by different substances, more especially as the bright yellow line of sodium was often seen in flames in which that metal was not supposed to be present. Professor Swan, having ascertained that the presence of the 2,500,000th part of a grain of sodium in a flame was sufficient to produce it, considered himself justified in asserting, in 1856, that this line was always to be taken as an indication of the presence of sodium in larger or smaller quantity.

But the greatest advance in spectral analysis was made by Bunsen and Kirchhoff, who, by means of a four-prism spectroscope, obtained accurate observations of the positions of the bright lines in the spectra of a great number of substances, as well as of the dark lines in the solar spectrum, and called attention to the identity of several of the latter with several of the former. Since the publication of their researches, the spectroscope has come into general use among chemists, and has already led to the discovery of four new metals, cæsium, rubidium, thallium, and indium.

1057. Reversal of Bright Lines. Analysis of the Sun's Atmosphere.

[1] These appear to be merely examples of the dissociation of the elements of a chemical compound at high temperatures.

—It may seem surprising that, while incandescent solids and liquids are found to give continuous spectra containing rays of all refrangibilities, the solar spectrum is interrupted by dark lines indicating the absence or relative feebleness of certain elementary rays. It seems natural to suppose that the deficient rays have been removed by selective absorption, and this conjecture was thrown out long since. But where and how is this absorption produced? These questions have now received an answer which appears completely satisfactory.

According to the theory of exchanges, which has been explained in connection with the radiation of heat (§ 464, 483), every substance which emits certain kinds of rays to the exclusion of others, absorbs the same kind which it emits; and when its temperature is the same in the two cases compared, its emissive and absorbing power are precisely equal for any one elementary ray.

When an incandescent vapour, emitting only rays of certain definite refrangibilities, and therefore having a spectrum of bright lines, is interposed between the observer and a very bright source of light, giving a continuous spectrum, the vapour allows no rays of its own peculiar kinds to pass; so that the light which actually comes to the observer consists of transmitted rays in which these particular kinds are wanting, together with the rays emitted by the vapour itself, these latter being of precisely the same kind as those which it has refused to transmit. It depends on the relative brightness of the two sources whether these particular rays shall be on the whole in excess or defect as compared with the rest. If the two sources are at all comparable in brightness, these rays will be greatly in excess, inasmuch as they constitute the whole light of the one, and only a minute fraction of the light of the other; but the light of the electric lamp, or of the lime-light, is usually found sufficiently powerful to produce the contrary effect; so that if, for example, a spirit-lamp with salted wick is interposed between the slit of a spectroscope and the electric light, the bright yellow line due to the sodium appears black by contrast with the much brighter back-ground which belongs to the continuous spectrum of the charcoal points. By employing only some 10 or 15 cells, a light may be obtained, the yellow portion of which, as seen in a one-prism spectroscope, is sensibly equal in brightness to the yellow line of the sodium flame, so that this line can no longer be separately detected, and the appearance is the same whether the sodium flame be interposed or removed.

The dark lines in the solar spectrum would therefore be accounted

for by supposing that the principal portion of the sun's light comes from an inner stratum which gives a continuous spectrum, and that a layer external to this contains vapours which absorb particular rays, and thus produce the dark lines. The stratum which gives the continuous spectrum might be solid, liquid, or even gaseous, for the experiments of Frankland and Lockyer have shown that, as the pressure of a gas is increased, its bright lines broaden out into bands, and that the bands at length become so wide as to join each other and form a continuous spectrum.[1]

Hydrogen, sodium, calcium, barium, magnesium, zinc, iron, chromium, cobalt, nickel, copper, and manganese have all been proved to exist in the sun by the accurate identity of position of their bright lines with certain dark lines in the sun's spectrum.

The strong line D, which in a good instrument is seen to consist of two lines near together, is due to sodium; and the lines C and F are due to hydrogen. No less than 450 of the solar dark lines have been identified with bright lines of iron.

1058. Telespectroscope. Solar Prominences.—For astronomical investigations, the spectroscope is usually fitted to a telescope, and takes the place of the eye-piece, the plane of the slit being placed in the principal focus of the object-glass, so that the image is thrown upon it, and the light which enters the slit is the light which forms one strip (so to speak) of the image, and which therefore comes from one strip of the object. A telescope thus equipped is called a telespectroscope. Extremely interesting results have been obtained by thus subjecting to examination a strip of the sun's edge, the strip being sometimes tangential to the sun's disc, and sometimes radial. When the former arrangement is adopted, the appearance presented is that depicted in No. 2, Plate III., consisting of a few bright lines scattered through a back-ground of the ordinary solar spectrum. The bright lines are due to an outer layer called the *chromosphere*, which is thus proved to be vaporous. The ordinary solar spectrum which accompanies it, is due to that part of the sun from which most of our light is derived. This part is called the *photosphere*, and if not solid or liquid, it must consist of vapour so highly compressed that its properties approximate to those of a liquid.

When the slit is placed radially, in such a position that only a

[1] The gradual transition from a spectrum of bright lines to a continuous spectrum may be held to be an illustration of the continuous transition which can be effected from the condition of ordinary gas to that of ordinary liquid (§ 380).

small portion of its length receives light from the body of the sun, the spectra of the photosphere and chromosphere are seen in immediate contiguity, and the bright lines in the latter (notably those of hydrogen, No. 14, Plate III.) are observed to form continuations of some of the dark lines of the former.

The chromosphere is so much less bright than the photosphere, that, until a few years since, its existence was never revealed except during total eclipses of the sun, when projecting portions of it were seen extending beyond the dark body of the moon. The spectrum of these projecting portions, which have been variously called "prominences," "red flames," and "rose-coloured protuberances," was first observed during the "Indian eclipse" of 1868, and was found to consist of bright lines, including those of hydrogen. From their excessive brightness, M. Janssen, who was one of the observers, expressed confidence that he should be able to see them in full sunshine; and the same idea had been already conceived and published by Mr. Lockyer. The expectation was shortly afterwards realized by both these observers, and the chromosphere has ever since been an object of frequent observation. The visibility of the chromosphere lines in full sunshine, depends upon the principle that, while a continuous spectrum is extended, and therefore made fainter, by increased dispersion, a bright line in a spectrum is not sensibly broadened, and therefore loses very little of its intrinsic brightness (§ 1061). Very high dispersion is necessary for this purpose.

Still more recently, by opening the slit to about the average width of the prominence-region, as measured on the image of the sun which is thrown on the slit, it has been found possible to see the whole of an average-sized prominence at one view. This will be understood by remembering that a bright line as seen in a spectrum is a monochromatic image of the illuminated portion of the slit, or when a telespectroscope is used, as in the present case, it is a monochromatic image of one strip of the image formed by the object-glass, namely, that strip which coincides with the slit. If this strip then contains a prominence in which the elementary rays C and F (No. 2, Plate III.) are much stronger than in the rest of the strip, a red image of the prominence will be seen in the part of the spectrum corresponding to the line C, and a blue image in the place corresponding to the line F. This method of observation requires greater dispersion than is necessary for the mere detection of the chromosphere lines; the dispersion required for enabling a bright-line spectrum to predomi-

nate over a continuous spectrum being always nearly proportional to the width of the slit (§ 1061).

Of the nebulæ, it is well known that some have been resolved by powerful telescopes into clusters of stars, while others have as yet proved irresolvable. Huggins has found that the former class of nebulæ give spectra of the same general character as the sun and the fixed stars, but that some of the latter class give spectra of bright lines, indicating that their constitution is gaseous.

1059. Displacement of Lines consequent on Celestial Motions.— According to the undulatory theory of light, which is now universally accepted, the fundamental difference between the different rays which compose the complete spectrum, is a difference of wave-frequency, and, as connected with this, a difference of wave-length in any given medium, the rays of greatest wave-frequency or shortest wave-length being the most refrangible.

Doppler first called attention to the change of refrangibility which must be expected to ensue from the mutual approach or recess of the observer and the source of light, the expectation being grounded on reasoning which we have explained in connection with acoustics (§ 898).

Doppler adduced this principle to explain the colours of the fixed stars, a purpose to which it is quite inadequate; but it has rendered very important service in connection with spectroscopic research. Displacement of a line towards the more refrangible end of the spectrum, indicates approach, displacement in the opposite direction indicates recess, and the velocity of approach or recess admits of being calculated from the observed displacement.

When the slit of the spectroscope crosses a spot on the sun's disc, the dark lines lose their straightness in this part, and are bent, sometimes to one side, sometimes to the other. These appearances clearly indicate uprush and downrush of gases in the sun's atmosphere in the region occupied by the spot.

Huggins detected a displacement of the F line towards the red end, in the spectrum of Sirius, as compared with the spectrum of the sun or of hydrogen. The displacement is so small as only to admit of measurement by very powerful instrumental appliances; but, small as it is, calculation shows that it indicates a motion of recess at the rate of about 30 miles per second.[1]

[1] The observed displacement corresponded to recess at the rate of 41·4 miles per second; but 12·0 of this must be deducted for the motion of the earth in its orbit at the season of

1060. Spectra of Artificial Lights.—The spectra of the artificial lights in ordinary use (including gas, oil-lamps, and candles) differ from the solar spectrum in the relative brightness of the different colours, as well as in the entire absence of dark lines. They are comparatively strong in red and green, but weak in blue; hence all colours which contain much blue in their composition appear to disadvantage by gas-light.

It is possible to find artificial lights whose spectra are of a completely different character. The salts of strontium, for example, give red light, composed of the ingredients represented in spectrum No. 10, Plate III., and those of sodium yellow light (No. 3, Plate III.). If a room is illuminated by a sodium flame (for example, by a spirit-lamp with salt sprinkled on the wick), all objects in the room will appear of a uniform colour (that of the flame itself), differing only in brightness, those which contain no yellow in their spectrum as seen by day-light being changed to black. The human countenance and hands assume a ghastly hue, and the lips are no longer red.

A similar phenomenon is observed when a coloured body is held in different parts of the solar spectrum in a dark room, so as to be illuminated by different kinds of monochromatic light. The object either appears of the same colour as the light which falls upon it, or else it refuses to reflect this light and appears black. Hence a screen for exhibiting the spectrum should be white.

1061. Brightness and Purity.—The laws which determine the brightness of images generally, and which have been expounded at some length in the preceding chapter, may be applied to the spectroscope. We shall, in the first instance, neglect the loss of light by reflection and imperfect transmission.

Let Δ denote the *prismatic dispersion*, as measured by the angular separation of two specified monochromatic images when the naked eye is applied to the last prism, the observing telescope being removed. Then, putting m for the linear magnifying power of the

the year when the observation was made. The remainder, 29·4, was therefore the rate at which the distance between the sun and Sirius was increasing.

In a more recent paper Dr. Huggins gave the results of observations with more powerful instrumental appliances. The recess of Sirius was found to be only 20 miles per second. Arcturus was found to be approaching at the rate of 50 miles per second. Community of motion was established in certain sets of stars; and the belief previously held by astronomers, as to the direction in which the solar system is moving with respect to the stars as a whole, was fully confirmed.

telescope, $m\Delta$ is the angular separation observed when the eye is applied to the telescope. We shall call $m\Delta$ the *total dispersion*.

Let θ denote the angle which the breadth of the slit subtends at the centre of the collimating lens, and which is measured by $\frac{\text{breadth of slit}}{\text{focal length of lens}}$. Then θ is also the apparent breadth of any absolutely monochromatic image of the slit, formed by rays of minimum deviation, as seen by an eye applied either to the first prism, the last prism, or any one of the train of prisms. The change produced in a pencil of monochromatic rays by transmission through a prism at minimum deviation, is in fact simply a change of direction, without any change of mutual inclination; and thus neither brightness nor apparent size is at all affected. In ordinary cases, the bright lines of a spectrum may be regarded as monochromatic, and their apparent breadth, as seen without the telescope, is sensibly equal to θ. Strictly speaking, the effect of prismatic dispersion in actual cases, is to increase the apparent breadth by a small quantity, which, if all the prisms are alike, is proportional to the number of prisms; but the increase is usually too small to be sensible.

Let I denote the intrinsic brightness of the source as regards any one of its (approximately) monochromatic constituents; in other words, the brightness which the source would have if deprived of all its light except that which goes to form a particular bright line. Then, still neglecting the light stopped by the instrument, the brightness of this line as seen without the aid of the telescope will be I; and as seen in the telescope it will either be equal to or less than this, according to the magnifying power of the telescope and the effective aperture of the object-glass (§ 1038). If the breadth of the slit be halved, the breadth of the bright line will be halved, and its brightness will be unchanged. These conclusions remain true so long as the bright line can be regarded as practically monochromatic.

The brightness of any part of a *continuous* spectrum follows a very different law. It varies directly as the width of the slit, and inversely as the prismatic dispersion. Its value without the observing telescope, or its maximum value with a telescope, is $\frac{\theta}{\Delta}i$, where i is a coefficient depending only on the source.

The *purity* of any part of a continuous spectrum is properly measured by the ratio of the *distance between two specified monochromatic images* to the *breadth of either*, the distance in question being measured from the centre of one to the centre of the other.

This ratio is unaffected by the employment of an observing telescope, and is $\frac{\Delta}{\theta}$.

The ratio of the brightness of a bright line to that of the adjacent portion of a continuous spectrum forming its back-ground, is $\frac{\Delta I}{\theta i}$, assuming the line to be so nearly monochromatic that the increase of its breadth produced by the dispersion of the prisms is an insignificant fraction of its whole breadth. As we widen the slit, and so increase θ, we must increase Δ in the same ratio, if we wish to preserve the same ratio of brightness. As $\frac{\Delta}{\theta}$ is increased indefinitely, the predominance of the bright lines does not increase indefinitely, but tends to a definite limit, namely, to the predominance which they would have in a perfectly pure spectrum of the given source.

The loss of light by reflection and imperfect transmission, increases with the number of surfaces of glass which are to be traversed; so that, with a long train of prisms and an observing telescope, the actual brightness will always be much less than the theoretical brightness as above computed.

The actual purity is always less than the theoretical purity, being greatly dependent on freedom from optical imperfections; and these can be much more completely avoided in lenses than in prisms. It is said that a single good prism, with a first-class collimator and telescope (as originally employed by Swan), gives a spectrum much more free from blurring than the modern multiprism spectroscopes, when the total dispersion $m\Delta$ is the same in both the cases compared.

1062. Chromatic Aberration.—The unequal refrangibility of the different elementary rays is a source of grave inconvenience in connection with lenses. The focal length of a lens depends upon its index of refraction, which of course increases with refrangibility, the focal length being shortest for the most refrangible rays. Thus a lens of uniform material will not form a single white image of a white object, but a series of images, of all the colours of the spectrum, arranged at different distances, the violet images being nearest, and the red most remote. If we place a screen anywhere in the series of images, it can only be in the right position for one colour. Every other colour will give a blurred image, and the superposition of them all produces the image actually formed on the screen. If the object be a uniform white spot on a black ground, its image on the screen

The inability of a single lens to bring rays to a focus is called Chromatic aberration.

will consist of white in its central parts, gradually merging into a coloured fringe at its edge. Sharpness of outline is thus rendered impossible, and nothing better can be done than to place the screen at the focal distance corresponding to the brightest part of the spectrum. Similar indistinctness will attach to images observed in mid-air, whether directly or by means of another lens. This source of confusion is called *chromatic aberration.*

1063. Possibility of Achromatism.—In order to ascertain whether it was possible to remedy this evil by combining lenses of two different materials, Newton made some trials with a compound prism composed of glass and water (the latter containing a little sugar of lead), and he found that it was not possible, by any arrangement of these two substances, to produce deviation of the transmitted light without separation into its component colours. Unfortunately he did not extend his trials to other substances, but concluded at once that an *achromatic* prism (and hence also an achromatic lens) was an impossibility; and this conclusion was for a long time accepted as indisputable. Mr. Hall, a gentleman of Worcestershire, was the first to show that it was erroneous, and is said to have constructed some achromatic telescopes; but the important fact thus discovered did not become generally known till it was rediscovered by Dollond, an eminent London optician, in whose hands the manufacture of achromatic instruments attained great perfection.

1064. Conditions of Achromatism.—The conditions necessary for achromatism are easily explained. The angular separation between the brightest red and the brightest violet ray transmitted through a prism is called the *dispersion* of the prism, and is evidently the difference of the deviations of these rays. These deviations, for the position of minimum deviation of a prism of small refracting angle A, are $(\mu'-1)$ A and $(\mu''-1)$ A, μ' and μ'' denoting the indices of refraction for the two rays considered (§ 1004, equation (1)) and their difference is $(\mu''-\mu')$ A. This difference is always small in comparison with either of the deviations whose difference it is, and its ratio to either of them, or more accurately its ratio to the value of $(\mu-1)$ A for the brightest part of the spectrum, is called the *dispersive power* of the substance. As the common factor A may be omitted, the formula for the dispersive power is evidently $\frac{\mu''-\mu'}{\mu-1}$.

If this ratio were the same for all substances, as Newton supposed, achromatism would be impossible; but in fact its value varies greatly,

and is greater for flint than for crown glass. If two prisms of these substances, of small refracting angles, be combined into one, with their edges turned opposite ways, they will achromatize one another if $(\mu'' - \mu')$ A, or the product of deviation by dispersive power, is the same for both. As the deviations can be made to have any ratio we please by altering the angles of the prisms, the condition is evidently possible.

The deviation which a simple ray undergoes in traversing a lens, at a distance x from the axis, is $\frac{x}{f}$, f denoting the focal length of the lens (§ 1004), and the separation of the red and violet constituents of a compound ray is the product of this deviation by the dispersive power of the material. If a convex and concave lens are combined, fitting closely together, the deviations which they produce in a ray traversing both, are in opposite directions, and so also are the dispersions. If we may regard x as having the same value for both (a supposition which amounts to neglecting the thicknesses of the lenses in comparison with their focal lengths) the condition of no resultant dispersion is that

$$\text{dispersive power} \times \frac{1}{f}$$ a Same for both

has the same value for both lenses. *Their focal lengths* must therefore be *directly as the dispersive powers of their materials.* These latter are about ·033 for crown and ·052 for flint glass. A converging achromatic lens usually consists of a double convex lens of crown fitted to a diverging meniscus of flint. In every achromatic combination of two pieces, the direction of resultant *deviation* is that due to the piece of smaller dispersive power.

The definition above given of dispersive power is rather loose. To make it accurate, we must specify, by reference to the "fixed lines," the precise positions of the two rays whose separation we consider.

Since the distances between the fixed lines have different proportions for crown and flint glass, achromatism of the whole spectrum is impossible. With two pieces it is possible to unite any two selected rays, with three pieces any three selected rays, and so on. It is considered a sign of good achromatism when no colours can be brought into view by bad focussing except purple and green.

1065. Achromatic Eye-pieces.—The eye-pieces of microscopes and astronomical telescopes, usually consist of two lenses of the same kind of glass, so arranged as to counteract, to some extent, the spherical

and chromatic aberrations of the object-glass. The *positive* eye-piece, invented by Ramsden, is suited for observation with cross-wires or micrometers; the *negative* eye-piece, invented by Huygens, is not adapted for purposes of measurement, but is preferred when distinct vision is the sole requisite. These eye-pieces are commonly called achromatic, but their achromatism is in a manner spurious. It consists not in bringing the red and violet images into true coincidence, but merely in causing one to cover the other as seen from the position occupied by the observer's eye.

In the best opera-glasses (§ 1033), the eye-piece, as well as the object-glass, is composed of lenses of flint and crown so combined as to be achromatic in the more proper sense of the word.

1066. Rainbow.—The unequal refrangibility of the different elementary rays furnishes a complete explanation of the ordinary phenomena of rainbows. The explanation was first given by Newton, who confirmed it by actual measurement.

It is well known that rainbows are seen when the sun is shining on drops of water. Sometimes one bow is seen, sometimes two, each of them presenting colours resembling those of the solar spectrum. When there is only one bow, the red arch is above and the violet below. When there is a second bow, it is at some distance outside of this, has the colours in reverse order, and is usually less bright.

Rainbows are often observed in the spray of cascades and fountains, when the sun is shining.

In every case, a line joining the observer to the sun is the axis of the bow or bows; that is to say, all parts of the length of the bow are at the same angular distance from the sun.

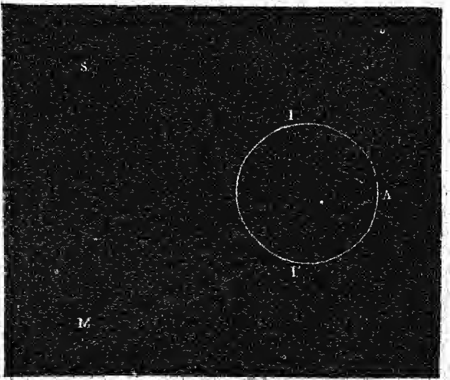


Fig. 765.—Production of Primary Bow.

The formation of the primary bow is illustrated by Fig. 765. A ray of solar light, falling on a spherical drop of water, in the direction SI, is refracted at I, then reflected internally from the back of the drop, and again refracted into the air in the

direction I′M. If we take different points of incidence, we shall obtain different directions of emergence, so that the whole light which emerges from the drop after undergoing, as in the figure, two refractions and one reflection, forms a widely-divergent pencil. Some portions of this pencil, however, contain very little light. This is especially the case with those rays which, having been incident nearly normally, are returned almost directly back, and also with those which were almost tangential at incidence. The greatest condensation, as regards any particular species of elementary ray, occurs at that part of the emergent pencil which has undergone *minimum deviation*. It is by means of rays which have undergone this minimum deviation, that the observer sees the corresponding colour in the bow; and the deviation which they have undergone is evidently equal to the angular distance of this part of the bow from the sun.

The minimum deviation will be greatest for those rays which are most refrangible. If the figure, for example, be supposed to represent the circumstances of minimum deviation for violet, we shall obtain smaller deviation in the case of red, even by giving the angle I A I′ the same value which it has in the case of minimum deviation for violet, and still more when we give it the value which corresponds to the minimum deviation of red. The most refrangible colours are accordingly seen furthest from the sun. The effect of the rays which undergo other than minimum deviation, is to produce a border of white light on the side remote from the sun; that is to say, on the inner edge of the bow.[1]

[1] When the drops are very uniform in size, a series of faint *supernumerary bows*, alternately purple and green, is sometimes seen beneath the primary bow. These bows are produced by the mutual interference of rays which have undergone other than minimum deviation, and the interference arises in the following way. Any two parallel directions of emergence, for rays of a given refrangibility, correspond in general to two different points of incidence on any given drop, one of the two incident rays being more nearly normal, and the other more nearly tangential to the drop than the ray of minimum deviation. These two rays have pursued dissimilar paths in the drop, and are in different phases when they reach the observer's eye. The difference of phase may amount to one, two, three, or more exact wave-lengths, and thus one, two, three, or more supernumerary bows may be formed. The distances between the supernumerary bows will be greater as the drops of water are smaller. This explanation is due to Dr. Thomas Young.

A more complete theory, in which diffraction is taken into account, is given by Airy in the *Cambridge Transactions* for 1838; and the volume for the following year contains an experimental verification by Miller. It appears from this theory that the maximum of intensity is less sharply marked than the ordinary theory would indicate, and does not correspond to the geometrical minimum of deviation, but to a deviation sensibly greater. Also that the region of sensible illumination extends beyond this geometrical minimum and shades off gradually.

The condensation which accompanies minimum deviation, is merely a particular case of the general mathematical law that magnitudes remain nearly constant in the neighbourhood of a maximum or minimum value. The rays which compose a small parallel pencil S I incident at and around the precise point which corresponds to minimum deviation, will thus have deviations which may be regarded as equal, and will accordingly remain sensibly parallel at emergence. A parallel pencil incident on any other part of the drop, will be divergent at emergence.

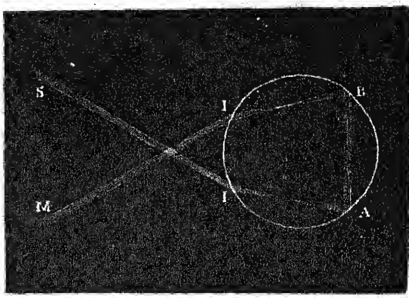


Fig. 766.—Production of Secondary Bow.

The indices of refraction for red and violet rays from air into water are respectively $\frac{108}{81}$ and $\frac{109}{81}$, and calculation shows that the distances from the centre of the sun to the parts of the bow in which these colours are strongest should be the supplements of 42° 2′ and 40° 17′ respectively. These results agree with observation. The angles 42° 2′ and 40° 17′ are the distances from the *anti-solar point,* which is always the centre of the bow.

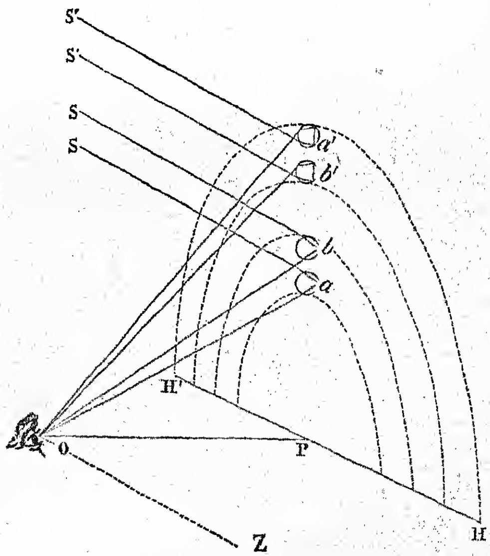


Fig. 767.—Relative Positions.

The rays which form the secondary bow have undergone two internal reflections, as represented in Fig. 766, and here again a special concentration occurs in the direction of minimum deviation. This deviation is greater than 180° and is greatest for the most refrangible rays. The distance of the arc thus formed from the sun's centre, is 360° *minus* the deviation, and is accord-

ingly least for the most refrangible rays. Thus the violet arc is nearest the sun, and the red furthest from it, in the secondary bow.

Some idea of the relative situations of the eye, the sun, and the drops of water in which the two bows are formed, may be obtained from an inspection of Fig. 767.

CHAPTER LXXIII.

COLOUR.

1067. Colour as a Property of Opaque Bodies.—A body which reflects (by irregular reflection) all the rays of the spectrum in equal proportion, will appear of the same colour as the light which falls upon it; that is to say, in ordinary cases, white or gray. But the majority of bodies reflect some rays in larger proportion than others, and are therefore coloured, their colour being that which arises from the mixture of the rays which they reflect. A body reflecting no light would be perfectly black. Practically, white, gray, and black differ only in brightness. A piece of white paper in shadow appears gray, and in stronger shadow black.

1068. Colour of Transparent Bodies.—A transparent body, seen by transmitted light, is coloured, if it is more transparent to some rays than to others, its colour being that which results from mixing the transmitted rays. No new ingredient is added by transmission, but certain ingredients are more or less completely stopped out.

Some transparent substances appear of very different colours according to their thickness. A solution of chloride of chromium, for example, appears green when a thin layer of it is examined, while a greater thickness of it presents the appearance of reddish brown. In such cases, different kinds of rays successively disappear by selective absorption, and the transmitted light, being always the sum of the rays which remain unabsorbed, is accordingly of different composition according to the thickness.

When two pieces of coloured glass are placed one behind the other, the light which passes through both has undergone a double process of selective absorption, and therefore consists mainly of those rays which are abundantly transmitted by both glasses; or to speak broadly, the colour which we see in looking through the combination

is not the sum of the colours of the two glasses, but their common part. Accordingly, if we combine a piece of ordinary red glass, transmitting light which consists almost entirely of red rays, with a piece of ordinary green glass, which transmits hardly any red, the combination will be almost black. The light transmitted through two glasses of different colour and of the same depth of tint, is always less than would be transmitted by a double thickness of either; and the colour of the transmitted light is in most cases a colour which occupies in the spectrum an intermediate place between the two given colours. Thus, if the two glasses are yellow and blue, the transmitted light will, in most cases, be green, since most natural yellows and blues when analysed by a prism show a large quantity of green in their composition. Similar effects are obtained by mixing coloured liquids.

1069. Colours of Mixed Powders.—"In a coloured powder, each particle is to be regarded as a small transparent body which colours light by selective absorption. It is true that powdered pigments when taken in bulk are extremely opaque. Nevertheless, whenever we have the opportunity of seeing these substances in compact and homogeneous pieces before they have been reduced to powder, we find them transparent, at least when in thin slices. Cinnabar, chromate of lead, verdigris, and cobalt glass are examples in point.

"When light falls on a powder thus composed of transparent particles, a small part is reflected at the upper surface; the rest penetrates, and undergoes partial reflection at some of the surfaces of separation between the particles. A single plate of uncoloured glass reflects $\frac{1}{25}$ of normally incident light; two plates $\frac{1}{13}$, and a large number nearly the whole. In the powder of such glass, we must accordingly conclude that only about $\frac{1}{25}$ of normally incident light is reflected from the first surface, and that all the rest of the light which gives the powder its whiteness comes from deeper layers. It must be the same with the light reflected from blue glass; and in coloured powders generally only a very small part of the light which they reflect comes from the first surface; it nearly all comes from beneath. The light reflected from the first surface is white, except when the reflection is metallic. That which comes from below is coloured, and so much the more deeply the further it has penetrated. This is the reason why coarse powder of a given material is more deeply coloured than fine, for the quantity of light returned at each successive reflection depends only on the number of reflections and not on the

thickness of the particles. If these are large, the light must penetrate so much the deeper in order to undergo a given number of reflections, and will therefore be the more deeply coloured.

"The reflection at the surfaces of the particles is weakened if we interpose between them, in the place of air, a fluid whose index of refraction more nearly approaches their own. Thus powders and pigments are usually rendered darker by wetting them with water, and still more with the more highly refracting liquid, oil.

"If the colours of powders depended only on light reflected from their first surfaces, the light reflected from a mixed powder would be the sum of the lights reflected from the surfaces of both. But most of the light, in fact, comes from deeper layers, and having had to traverse particles of both powders, must consist of those rays which are able to traverse both. The resultant colour therefore, as in the case of superposed glass plates, depends not on addition but rather on subtraction. Hence it is that a mixture of two pigments is usually much more sombre than the pigments themselves, if these are very unlike in the average refrangibility of the light which they reflect. Vermilion and ultramarine, for example, give a black-gray (showing scarcely a trace of purple, which would be the colour obtained by a true mixture of lights), each of these pigments being in fact nearly opaque to the light of the other."[1]

1070. Mixtures of Colours.—By the colour resulting from the mixture of two lights, we mean the colour which is seen when they both fall on the same part of the retina. Propositions regarding mixtures of colours are merely subjective. The only objective differences of colour are differences of refrangibility, or if traced to their source, differences of wave-frequency. All the colours in a pure spectrum are objectively simple, each having its own definite period of vibration by which it is distinguished from all others. But whereas, in acoustics, the quality of a sound as it affects the ear varies with every change in its composition, in colour, on the other hand, very different compositions may produce precisely the same visual impression. Every colour that we see in nature can be exactly imitated by an infinite variety of different combinations of elementary rays.

To take, for example, the case of white. Ordinary white light consists of all the colours of the spectrum combined; but any one of the elementary colours, from the extreme red to a certain point in yellowish green, can be combined with another elementary colour

[1] Translated from Helmholtz's *Physiological Optics*, § 20.

on the other side of green in such proportion as to yield a perfect imitation of ordinary white. The prism would instantly reveal the differences, but to the naked eye all these whites are completely undistinguishable one from another.

1071. Methods of Mixing Colours.—The following are some of the best methods of mixing colours (that is coloured lights):—

1. By combining reflected and transmitted light; for example, by looking at one colour through a piece of glass, while another colour is seen by reflection from the near side of the glass. The lower sash of a window, when opened far enough to allow an arm to be put through, answers well for this purpose. The brighter of the two coloured objects employed should be held inside the window, and seen by reflection; the second object should then be held outside in such a position as to be seen in coincidence with the image of the first. As the quantity of reflected light increases with the angle of incidence, the two colours may be mixed in various proportions by shifting the position of the eye. This method is not however adapted to quantitative comparison, and can scarcely be employed for combining more than two colours.

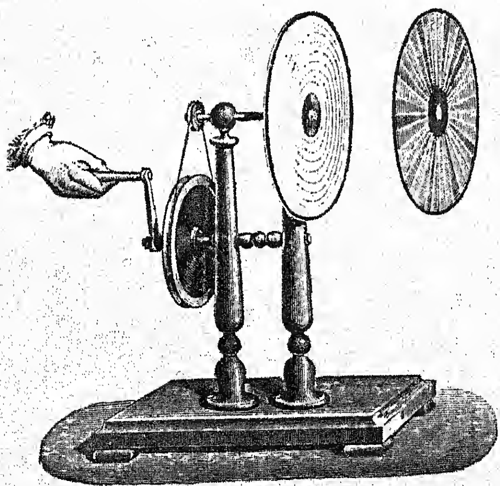


Fig. 768.—Rotating Disc.

2. By employing a rotating disc (Fig. 768) composed of differently coloured sectors. If the disc be made to revolve rapidly, the sectors will not be separately visible, but their colours will appear blended into one on account of the persistence of visual impressions. The proportions can be varied by varying the sizes of the sectors. Coloured discs of paper, each having a radial slit, are very convenient for this purpose, as any moderate number of such discs can be combined, and the sizes of the sectors exhibited can be varied at pleasure.

The mixed colour obtained by a rotating disc is to be regarded as

a *mean* of the colours of the several sectors—a mean in which each of these colours is assigned a weight proportional to the size of its sector. Thus, if the 360 degrees which compose the entire disc consist of 100° of red paper, 100° of green, and 160° of blue, the intensity of the light received from the red when the disc is rotating will be only $\frac{10}{36}$ of that which would be received from the red sector when seen at rest; and the total effect on the retina is represented by $\frac{10}{36}$ of the intensity of the red, *plus* $\frac{10}{36}$ of the intensity of the green, *plus* $\frac{16}{36}$ of the intensity of the blue; or if we denote the colours of the sectors by their initial letters, the effect may be symbolized by the formula $\frac{10\,R+10\,G+16\,B}{36}$. Denoting the resultant colour by C, we have the symbolic equation

$$10\,R+10\,G+16\,B=36\,C;$$

and the resultant colour may be called the mean of 10 parts of red, 10 of green, and 16 of blue. Colour-equations, such as the above, are frequently employed, and may be combined by the same rules as ordinary equations.

3. By causing two or more spectra to overlap. We thus obtain mixtures which are the *sums* of the overlapping colours.

If, in the experiment of § 1048, we employ, instead of a single straight slit, a pair of slits meeting at an angle, so as to form either an X or a V, we shall obtain mixtures of all the simple colours two and two, since the coloured images of one of the slits will cross those of the other. The display of colours thus obtained upon a screen is exquisitely beautiful, and if the eye is placed at any point of the image (for example, by looking through a hole in the screen), the prism will be seen filled with the colour which falls on this point.

1072. Experiments of Helmholtz and Maxwell.—Helmholtz, in an excellent series of observations of mixtures of simple colours, employed a spectroscope with a V-shaped slit, the two strokes of the V being at right angles to one another; and by rotating the V he was able to diminish the breadth and increase the intensity of one of the two spectra, while producing an inverse change in the other. To isolate any part of the compound image formed by the two overlapping spectra, he drew his eye back from the eye-piece, so as to limit his view to a small portion of the field.

But the most effective apparatus for observing mixtures of simple colours is one devised by Professor Clerk Maxwell, by means of which any two or three colours of the spectrum can be combined in

any required proportions. In principle, this method is nearly equivalent to looking through the hole in the screen in the experiment above described.

Let P (Fig. 769) be a prism, in the position of minimum deviation; L a lens; E and R conjugate foci for rays of a particular refrangibility, say red; E and V conjugate foci for rays of another given refrangibility, say violet. If a slit is opened at R, an eye at E will receive only red rays, and will see the lens filled with red light. If this slit be closed, and a slit opened at V, the eye, still placed at E, will see the lens filled with violet light. If both slits be opened, it will see the lens filled with a uniform mixture of the two lights; and if a third slit be opened, between R and V, the lens will be seen filled with a mixture of three lights.

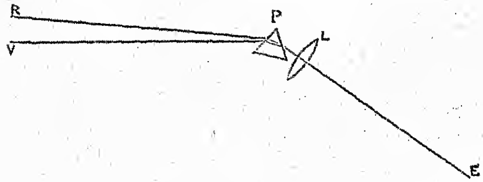


Fig. 769.—Principle of Maxwell's Colour-box.

Again, from the properties of conjugate foci, if a slit is opened at E, its spectral image will be formed at R V, the red part of it being at R, and the violet part at V.

The apparatus was inclosed in a box painted black within. There was a slit fixed in position at E, and a frame with three movable slits at R V. When it was desired to combine colours from three given parts of the spectrum, specified by reference to Fraunhofer's lines, the slit E was first turned towards the light, giving a real spectrum in the plane R V, in which Fraunhofer's lines were visible, and the three movable slits were set at the three specified parts of the spectrum. The box was then turned end for end, so that light was admitted (reflected from a large white screen placed in sunshine) at the movable slits, and the observer, looking in at the slit E, saw the resultant colour.

1073. Results of Experiment.—The following are some of the principal results of experiments on the mixture of coloured lights:—

1. Lights which appear precisely alike to the naked eye yield identical results in mixtures; or employing the term *similar* to express apparent identity as judged by the naked eye, *the sums of similar lights are themselves similar*. It is by reason of this physical fact, that colour-equations yield true results when combined according to the ordinary rules of elimination.

In the strict application of this rule, the same observer must be the judge of similarity in the different cases considered. For

2. Colours may be similar as seen by one observer, and dissimilar as seen by another; and in like manner, colours may be similar as seen through one coloured glass, and dissimilar as seen through another. The reason, in both cases, is that selective absorption depends upon real composition, which may be very different for two merely similar lights. Most eyes are found to exhibit selective absorption of a certain kind of elementary blue, which is accordingly weakened before reaching the retina.

3. Between any four colours, given in intensity as well as in kind, one colour-equation subsists; expressing the fact that, when we have the power of varying their intensities at pleasure, there is one definite way of making them yield a *match*, that is to say, a pair of similar colours. Any colour can therefore be completely specified by three numbers, expressing its relation to three arbitrarily selected colours. This is analogous to the theorem in statics that a force acting at a given point can be specified by three numbers denoting its components in three arbitrarily selected directions.

4. Between any five colours, given in intensity as well as in kind, a match can be made in one definite way by taking means;[1] for example, by mounting the colours on two rotating discs. If we had the power of illuminating one disc more strongly than the other in any required ratio, four colours would be theoretically sufficient; and we can, in fact, do what is nearly equivalent to this, by employing black as one of our five colours. Taking means of colours is analogous to finding centres of gravity. In following out the analogy, a colour given in kind merely must be represented by a material point given in position merely, and the intensity of the colour must be represented by the mass of the material point. The means of two given colours will be represented by points in the line joining two given points. The means of three given colours will be represented by points lying within the triangle formed by joining three given points, and the means of four given colours will be represented by points within a tetrahedron whose four corners are given. When we have five colours given, we have five points given, and of these generally no four will lie in one plane. Call them A, B, C, D, E.

[1] Propositions 4 and 5 are not really independent, but represent different aspects of one physical (or rather physiological) law.

Then if E lies within the tetrahedron A B C D, we can make the centre of gravity of A, B, C, and D coincide with E, and the colour E can be matched by a mean of the other four colours.

If E lies outside the tetrahedron, it must be situated at a point from which either one, two, or three faces are visible (the tetrahedron being regarded as opaque).

If only one face is visible, let it be B C D; then the point where the straight line E A cuts B C D is the match; for it is a mean of E and A, and is also a mean of B, C, and D.

If two faces are visible, let them be A C D and B C D; then the intersection of the edge C D with the plane E A B is the match.

If three faces are visible, let them be the three which meet at A; then A is the match, for it lies within the tetrahedron E B C D.

With six given colours, combined five at a time, six different matches can be made, and six colour-equations will thus be obtained, the consistency of which among themselves will be a test of the accuracy both of theory and observation, as only three of the six can be really independent. Experiments which have been conducted on this plan have given very consistent results.

1074. Cone of Colour.—All the results of mixing colours can be represented geometrically by means of a cone or pyramid within which all possible colours will have their definite places. The vertex will represent total blackness, or the complete absence of light; and colours situated on the same line passing through the vertex will differ only in intensity of light. Any cross-section of the cone will contain all colours, except so far as intensity is concerned, and the colours residing on its perimeter will be the colours of the spectrum ranged in order, with purple to fill up the interval between violet and red. It appears from Maxwell's experiments, that the true form of the cross-section is approximately triangular;[1] with red, green, and violet at the three corners. When all the colours have been assigned their proper places in the cone, a straight line joining any two of them passes through colours which are means of these two; and if two lines are drawn from the vertex to any two colours, the parallelogram constructed on these two lines will have at its further corner the colour which is the sum of these two colours. A certain axial line of the cone will contain

[1] The shape of the triangle is a mere matter of convenience, not involving any question of fact.

white or gray at all points of its length, and may be called the *line of white*.

It is convenient to distinguish three qualities of colour which may be called *hue, depth,* and *brightness*. *Brightness* or *intensity* of light is represented by distance from the vertex of the cone. *Depth* depends upon angular distance from the line of white, and is the same for all points on the same line through the vertex. *Paleness* or *lightness* is the opposite of depth, and is measured by angular nearness to the line of white. *Hue* or *tint* is that which is often *par excellence* termed colour. If we suppose a plane, containing the line of white, to revolve about this line as axis, it will pass successively through different tints; and in any one position it contains only two tints, which are separated from each other by the line of white, and are complementary.

Red is complementary to	Bluish green.
Orange " "	Sky blue.
Yellow " "	Violet blue.
Greenish yellow "	Violet.
Green " "	Pink.

Any two colours of complementary tint give white, when mixed in proper proportions; and any three colours can be mixed in such proportions as to yield white, if the triangle formed by joining them is pierced by the line of white.

Every colour in nature, except purple, is similar to a colour of the spectrum either pure or diluted with gray; and all purples are similar to mixtures of red and blue with or without dilution. Brown can be imitated by diluting orange with dark gray. The orange and yellow of the spectrum can themselves be imitated by adding together red and green.

1075. Three Primary Colour-sensations.—All authorities are now agreed in accepting the doctrine, first propounded by Dr. Thomas Young, that there are three elements of colour-sensation; or, in other words, three distinct physiological actions, which, by their various combinations, produce our various sensations of colour. Each is excitable by light of various wave-lengths lying within a wide range, but has a maximum of excitability for a particular wave-length, and is affected only to a slight degree by light of wave-length very different from this. The cone of colour is theoretically a triangular pyramid, having for its three edges the colours which correspond to these three wave-lengths; but it is probable that we cannot obtain

one of the three elementary colour-sensations quite free from admixture of the other two, and the edges of the pyramid are thus practically rounded off. One of these sensations is excited in its greatest purity by the green near Fraunhofer's line *b*, another by the extreme red, and the third by the extreme violet.

Helmholtz ascribes these three actions to three distinct sets of nerves, having their terminations in different parts of the thickness of the retina—a supposition which aids in accounting for the approximate achromatism of the eye, for the three sets of nerve-terminations may thus be at the proper distances for receiving distinct images of red, green, and violet respectively, the focal length of a lens being shorter for violet than for red.

Light of great intensity, whatever its composition, seems to produce a considerable excitement of all three elements of colour-sensation. If a spectroscope, for example, be directed first to the clouds and then to the sun, all parts of the spectrum appear much paler in the latter case than in the former.

The popular idea that red, yellow, and blue are the three primaries, is quite wrong as regards mixtures of lights or combinations of colour-sensations. The idea has arisen from facts observed in connection with the mixture of pigments and the transmission of light through coloured glasses. We have already pointed out the true interpretation of observations of this nature, and have only now to add that in attempting to construct a theory of the colours obtained by mixtures of pigments, the law of substitution of *similars* cannot be employed. Two pigments of *similar* colour will not in general give the same result in mixtures.

1076. Accidental Images.—If we look steadily at a bright stained-glass window, and then turn our eyes to a white wall, we see an image of the window with the colours changed into their complementaries. The explanation is that the nerves which have been strongly exercised in the perception of the bright colours have had their sensibility diminished, so that the balance of action which is necessary to the sensation of white no longer exists, but those elements of sensation which have not been weakened preponderate. The subjective appearances arising from this cause are called *negative accidental images*. Many well-known effects of contrast are similarly explained. White paper, when seen upon a background of any one colour, often appears tinged with the complementary colour; and stray beams of sunlight entering a room shaded with

yellow holland blinds, produce blue streaks when they fall upon a white tablecloth.

In some cases, especially when the object looked at is painfully bright, there is a *positive* accidental image; that is, one of the same colour as the object; and this is frequently followed by a negative image. A positive accidental image may be regarded as an extreme instance of the persistence of impressions.

1077. Colour-blindness.—What is called colour-blindness has been found, in every case which has been carefully investigated, to consist in the absence of the elementary sensation corresponding to red. To persons thus affected the solar spectrum appears to consist of two decidedly distinct colours, with white or gray at their place of junction, which is a little way on the less refrangible side of the line F. One of these two colours is doubtless nearly identical with the normal sensation of blue or violet. It attains its maximum about midway between F and G, and extends beyond G as far as the normally visible spectrum. The other colour extends a considerable distance into what to normal eyes is the red portion of the spectrum, attaining its maximum about midway between D and E, and becoming deeper and more faint till it vanishes at about the place where to normal eyes crimson begins. The scarlet of the spectrum is thus visible to the colour-blind, not as scarlet but as a deep dark colour, perhaps a kind of dark green, orange and yellow as brighter shades of the same colour, while bluish-green appears nearly white.

It is obvious from this account that what is called "colour-blindness" should rather be called *dichroic vision*, normal vision being distinctively designated as *trichroic*. To the dichroic eye any colour can be matched by a mixture of yellow and blue, and a match can be made between any three (instead of four) given colours. Objects which have the same colour to the trichroic eye have also the same colour to the dichroic eye.

1078. Colour and Musical Pitch.—As it is completely established that the difference between the colours of the spectrum is a difference of vibration-frequency, there is an obvious analogy between colour and musical pitch; but in almost all details the relations between colours are strikingly different from the relations between sounds.

The compass of visible colour, including the lavender rays which lie beyond the violet, and are perhaps visible not in themselves,

but by the fluorescence which they produce on the retina, is, according to Helmholtz, about an octave and a fourth; but if we exclude the lavender, it is almost exactly an octave. Attempts have been made to compare the successive colours of the spectrum with the notes of the gamut; but much forcing is necessary to bring out any trace of identity, and the gradual transitions which characterize the spectrum, and constitute a feature of its beauty, are in marked contrast to the transitions *per saltum* which are required in music.

CHAPTER LXXIV.

WAVE THEORY OF LIGHT.

1079. Principle of Huygens.[1]—The propagation of waves, whether of sound or light, is a propagation of energy. Each small portion of the medium experiences successive changes of state, involving changes in the forces which it exerts upon neighbouring portions. These changes of force produce changes of state in these neighbouring portions, or in such of them as lie on the forward side of the wave, and thus a disturbance existing at any one part is propagated onwards.

Let us denote by the name *wave-front* a continuous surface drawn through particles which have the same phase; then each wave-front advances with the velocity of light, and each of its points may be regarded as a secondary centre from which disturbances are continually propagated. This mode of regarding the propagation of light is due to Huygens, who derived from it the following principle, which lies at the root of all practical applications of the undulatory theory: *The disturbance at any point of a wave-front is the resultant* (given by the parallelogram of motions) *of the separate disturbances which the different portions of the same wave-front in any one of its earlier positions, would have occasioned if acting singly.* This principle involves the physical fact that rays of light are not affected by crossing one another; and its truth, which has been experimentally tested by a variety of consequences, must be taken as an indication that the amplitudes of luminiferous vibrations are infinitesimal in comparison with the wave-lengths. A similar law applies to the resultant of small disturbances generally, and is called by writers on dynamics the law of "superposition of small motions." It is analogous to the arithmetical principle that, when a and b are very small fractions, the product of $1+a$ and $1+b$ may be identified with

[1] For the spelling of this name see remarks by Lalande, *Mémoires de l'Académie*, 1773.

$1+a+b$; the term ab, which represents the mutual influence of two small changes, being negligible in comparison with the sum $a+b$ of the small changes themselves.

1080. Explanation of Rectilinear Propagation.—In a medium in which light travels with the same velocity in all parts and in all directions, the waves propagated from any point will be concentric spheres, having this point for centre; and the lines of propagation, in other words the rays of light, will be the radii of these spheres. It can in fact be shown that the only part of one of these waves which needs to be considered, in computing the resultant disturbance of an external point, is the part which lies directly between this external point and the centre of the sphere. The remainder of the wave-front can be divided into small parts, each of which, by the mutual interference of its own subdivisions, gives a resultant effect of zero at the given point. We express these properties by saying that *in a homogeneous and isotropic medium the wave-surface is a sphere, and the rays are normal to the wave-fronts.* This class of media includes gases, liquids, crystals of the cubic system, and well-annealed glass.

If a medium be homogeneous but not isotropic, disturbances emanating from a point in it will be propagated in waves which will retain their form unchanged as they expand in receding from their source, but this form will not generally be spherical. The rays of light in such a medium will be straight, proceeding directly from the centre of disturbance, and any one ray will cut all the wave-fronts at the same angle; but this angle will generally be different for different rays. In this case, as in the last, the disturbance produced at any point may be computed by merely taking into account that small portion of a wave-front which lies directly between the given point and the source,—in other words, which lies on or very near to the ray which traverses the given point.

A disturbance in such a medium usually gives rise to two sets of waves, having two distinct forms, and these remarks apply to each set separately.

The tendency of the different parts of a wave-front to propagate disturbances in other directions besides the single one to which such propagation is usually confined, is manifested in certain phenomena which are included under the general name of *diffraction*.

The only wave-fronts with which it is necessary to concern ourselves are those which belong to waves emanating from a single

point,—that is to say, either from a surface really very small, or from a surface which, by reason of its distance, subtends a very small solid angle at the parts of space considered.

1081. Application to Refraction.—When waves are propagated from one medium into another, the principle of Huygens leads to the following construction:—

Let A E (Fig. 770) represent a portion of the surface of separation between two media, and A B a portion of a wave-front in the first medium; both portions being small enough to be regarded as plane.

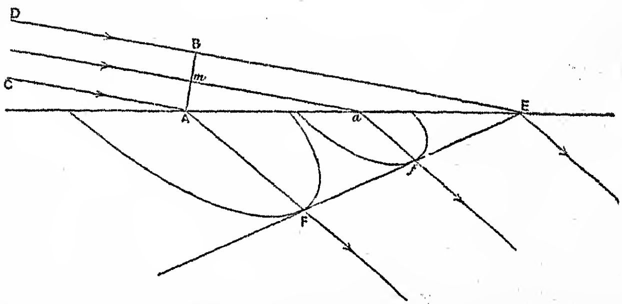


Fig. 770.—Huygens' Construction for Wave-front.

Then straight lines C A, D B E, normal to the wave-front, represent rays incident at A and E. From A as centre, describe a wave-surface, of such dimensions that light emanating from A would reach this surface in the same time in which light in air travels the distance B E, and draw a tangent plane (perpendicular to the plane of incidence) through E to this surface. Let F be the point of contact (which is not necessarily in the plane of incidence). Then the tangent plane E F is a wave-front in the second medium, and A F is a ray in the second medium; for it can be shown that disturbances propagated from all points in the wave-front A B will just have reached E F when the disturbance propagated from B has reached E. For example, a ray proceeding from m, the middle point of the line A B, will exhaust half the time in travelling to the middle point a of A E, and the remaining half in travelling through af, equal and parallel to half of A F.

When the wave-surfaces in both media are spherical, the planes of incidence and refraction A B E, A F E coincide, the angle B A E (Fig. 771) between the first wave-front and the surface of separation is the same as the angle between the normals to these surfaces, that

is to say, is the angle of incidence; and the angle A E F between the surface of separation and the second wave-front is the angle of refraction. The sine of the former is $\frac{BE}{EA}$, and the sine of the latter is $\frac{AF}{EA}$. The ratio $\frac{\sin i}{\sin r}$ is therefore $\frac{BE}{AF}$. But B E and A F are the distances travelled in the same time in the two media. Hence the sines of the angles of incidence and refraction are directly as the velocities of propagation of the incident and refracted light. The *relative index* of refraction from one medium into another is therefore the *ratio of the velocity of light in the first medium to its velocity in the second;* and *the absolute index of refraction of any medium is inversely as the velocity of light in that medium.*

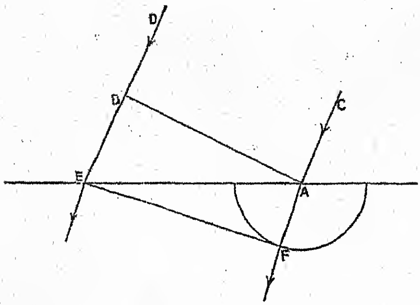


Fig. 771.—Wave-front in Ordinary Refraction.

1082. Application to Reflection.—The explanation of reflection is precisely similar. Let C A, D E (Fig. 772) be parallel rays incident at A and E; A B the wave-front. As the successive points of the wave-front arrive at the reflecting surface, hemispherical waves diverge from the points of incidence; and by the time that B reaches E, the wave from A will have diverged in all directions to a distance equal to B E. If then we describe in the plane of incidence a semicircle, with centre A and radius equal to B E, the tangent E F to this semicircle will be the wave-front of the reflected light, and A F will be the reflected ray corresponding to the incident ray C A. From the equality of the right-angled triangles A B E, E F A, it is evident that the angles of incidence and reflection are equal.

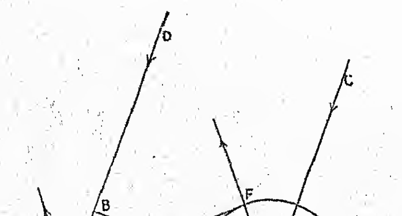


Fig. 772.—Wave-front in Reflection.

1083. Newtonian Explanation of Refraction.—In the Newtonian theory, the change of direction which a ray experiences at the bound-

ing surface of two media, is attributed to the preponderance of the attraction of the denser medium upon the particles of light. As the resultant force of this attraction is normal to the surface, the tangential component of velocity remains unchanged, and the normal component is increased or diminished according as the incidence is from rare to dense or from dense to rare. Let μ denote the relative index of refraction from rare to dense. Let v, v' be the velocities of light in the rarer and denser medium respectively, and i, i' the angles which the rays in the two media make with the normal. Then the tangential components of velocity in the two media are $v \sin i$, $v' \sin i'$ respectively, and these by the Newtonian theory are equal; whence $\frac{v'}{v}=\frac{\sin i}{\sin i'}=\mu$; whereas according to the undulatory theory $\frac{v'}{v}=\frac{1}{\mu}$. In the Newtonian theory, the velocity of light in any medium is directly as the absolute index of refraction of the medium; whereas, in the undulatory theory, the reverse rule holds.

The main design of Foucault's experiment with the rotating mirror (§ 942), in its original form, was to put these opposite conclusions to the test of direct experiment. For this purpose it was not necessary to determine the velocity of the rotating mirror, since it affected both the observed displacements alike. The two images were seen in the same field of view, and were easily distinguished by the greenness of the water-image. In every trial the water-image was more displaced than the air-image, indicating longer time and slower velocity; and the measurements taken were in complete accordance with the undulatory theory, while the Newtonian theory was conclusively disproved.

1084. Principle of Least Time.—The path by which light travels from one point to another is in the generality of cases that which occupies least time. For example, in ordinary cases of reflection (except from very concave[1] surfaces), if we select any two points, one on the incident and the other on the reflected ray, the sum of their distances from the point of incidence is less than the sum of their distances from any neighbouring point on the reflecting surface. In this case, since only one medium is concerned, distance is proportional to time. When a ray in air is refracted into water, if we select any two points,

[1] Suppose an ellipse described, having the two selected points for foci, and passing through the point of incidence. If the curvature of the reflecting surface in the plane of incidence is greater than the curvature of this ellipse, the length of the path is a maximum, if less, a minimum. This follows at once from the constancy of the sum of the focal distances in an ellipse.

one on the incident and the other on the refracted ray, and call their distances from any point of the refracting surface s, s' respectively, and the velocities of propagation in the two media v, v', then the sum of $\frac{s}{v}$ and $\frac{s'}{v'}$ is generally less when s and s' are measured to the point of incidence than when they are measured to any neighbouring point on the surface. $\frac{s}{v}$ is evidently the time of going from the first point to the refracting surface, and $\frac{s'}{v'}$ the time from the refracting surface to the second point.

The proposition as above enunciated admits of certain exceptions, the time being sometimes a maximum instead of a minimum. The really essential condition (which is fulfilled in both these opposite cases) is that all points on a small area surrounding the point of incidence give sensibly *the same time*. The component waves sent from all parts of this small area will be in the same phase, and will propagate a ray of light by their combined action.

When the two points considered are conjugate foci, and there is no aberration, this condition must be fulfilled by all the rays which pass through both; and the *time of travelling from one focus to the other is the same for all the rays*. Spherical waves diverging from one focus will, after incidence, become spherical waves converging to or diverging from its conjugate focus. An effect of this kind can be beautifully exhibited to the eye by means of an elliptic dish containing mercury. If agitation is produced at one focus of the ellipse by dipping a small rod into the liquid at this point, circular waves will be seen to converge towards the other focus. A circular dish exhibits a similar result somewhat imperfectly; waves diverging from a point near the centre will be seen to converge to a point symmetrically situated on the other side of the centre.

When the second point lies on a caustic surface formed by the reflection or refraction of rays emanating from the first point, all points on an area of sensible magnitude in the neighbourhood of the point of incidence would give sensibly the same time of travelling as the actual point of incidence, so that the light which traverses a point on a caustic may be regarded as coming from an area of sensible magnitude instead of (as in the case of points not on the caustic) an excessively small area. An eye placed at a point on a caustic will see this portion of the surface filled with light.

As the velocity of light is inversely proportional to the index of

refraction μ, the time of travelling a distance s with constant velocity may be represented by μs, and if a ray of light passes from one point to another by a crooked path, made up of straight lines $s_1, s_2, s_3, \ldots$ lying in media whose absolute indices are $\mu_1, \mu_2, \mu_3, \ldots$, the expression $\mu_1 s_1 + \mu_2 s_2 + \mu_3 s_3 + \ldots$ represents the time of passage. This expression, which may be called *the sum of such terms as* μs, must therefore fulfil the above condition; that is to say, the points of incidence on the surfaces of separation must be so situated that this sum either remains absolutely constant when small changes are supposed to be made in the positions of these points, or else retains that approximate constancy which is characteristic of maxima and minima. Conversely, all lines from a luminous point which fulfil this condition, will be paths of actual rays.

1085. Terrestrial Refraction.[1]—The atmosphere may be regarded as homogeneous when we confine our attention to small portions of it, and hence it is sensibly true, in ordinary experiments where no great distances are concerned, that rays of light in air are straight, just as it is true in the same limited sense that the surface of a liquid at rest is a horizontal plane. The surface of an ocean is not plane, but approximately spherical, its curvature being quite sensible in ordinary nautical observations, where the distance concerned is merely that of the visible sea-horizon; and a correction for curvature is in like manner required in observing levels on land. If the observer is standing on a perfectly level plain, and observing a distant object at precisely the same height as his eye above the plain, it will appear to be below his eye, for a horizontal *plane* through his eye will pass above it, since a perfectly level *plain* is not *plane*, but shares in the general curvature of the earth. It is easily proved that the apparent depression due to this cause is half the angle between the verticals at the positions of the observer and of the object observed. But experience has shown that this apparent depression is to a considerable extent modified by an opposite disturbing cause, called *terrestrial refraction.* When the atmosphere is in its normal condition, a ray of light from the object to the observer is not straight, but is slightly concave downwards.

This curvature of a nearly horizontal ray is not due to the curvature of the earth and of the layers of equal density in the earth's atmosphere, as is often erroneously supposed, but would still exist,

[1] For the leading idea which is developed in §§ 1085–1087, the Editor is indebted to suggestions from Professor James Thomson.

and with no sensible change in its amount, if the earth's surface were plane, and the directions of gravity everywhere parallel. It is due to the fact that light travels faster in the rarer air above than in the denser air below, so that time is saved by deviating slightly to the upper side of a straight course. The actual amount of curvature (as determined by surveying) is from $\frac{1}{2}$ to $\frac{1}{10}$ of the curvature of the earth; that is to say, the radius of curvature of the ray is from 2 to 10 times the earth's radius.

1086. Calculation of Curvature of Ray.—In order to calculate the radius of curvature from physical data, it is better to approach the subject from a somewhat different point of view.

The wave-fronts of a ray in air are perpendicular to the ray; and if the ray is nearly horizontal, its wave-fronts will be nearly vertical. If two of these wave-fronts are produced downwards until they meet, the distance of their intersection from the ray will be the radius of curvature. Let us consider two points on the same wave-front, one of them a foot above the other; then the upper one being in rarer air will be advancing faster than the lower one, and it is easily shown that the difference of their velocities is to the velocity of either, as 1 foot is to the radius of curvature.

Put ρ for the radius of curvature in feet, v and $v+\delta v$ for the two velocities, μ and $\mu-\delta\mu$ for the indices of refraction of the air at the two points. Then we have

$$\frac{1}{\rho} = \frac{\delta v}{v} = \frac{\delta \mu}{\mu} = \delta\mu \text{ nearly.} \tag{1}$$

Now it has been ascertained, by direct experiment, that the value of $\mu-1$ for air, within ordinary limits of density, is sensibly proportional to the density (even when the temperature varies), and is ·0002943 or $\frac{1}{3400}$ at the density corresponding to the pressure 760^{mm}, (at Paris) and temperature 0°C. The difference of density at the two points considered, supposing them both to be at the same temperature, will be to the density of either as 1 foot is to the "height of the homogeneous atmosphere" in feet, which call H (§ 211). Then $\frac{\delta\mu}{\mu-1}$ will be $\frac{1}{\mathrm{H}}$, and the value of $\frac{1}{\rho}$ in (1) may be written

$$\frac{1}{\rho} = \frac{\delta\mu}{\mu-1}(\mu-1) = \frac{1}{\mathrm{H}}(\mu-1) = \frac{1}{\mathrm{H}}\,\frac{1}{3400}. \tag{2}$$

Hence ρ is 3400 times the height of the homogeneous atmosphere. But this height is about 5 miles, or $\frac{1}{800}$ of the earth's radius. The

value of ρ is therefore about $4\frac{1}{4}$ radii of the earth. This is on the assumptions that the barometer is at 760^{mm}, the thermometer at 0° C., and that there is no change of temperature in ascending. If we depart from these assumptions, we have the following consequences:—

I. If the barometer is at any other height, the factor $\frac{1}{H}$ remains unaltered, and the other factor $\mu - 1$ varies directly as the pressure.

II. If the temperature is $t°$ Centigrade, H is changed in the direct ratio of $1+\alpha t$, α denoting the coefficient of expansion. The first factor $\frac{1}{H}$ is therefore changed in the inverse ratio of $1+\alpha t$. The second factor is changed in the same ratio. The curvature of the ray therefore varies inversely as $(1+\alpha t)^2$.

III. Suppose the temperature decreases upwards at the rate of $\frac{1}{n}$ of a degree Centigrade per foot. The expansion due to $\frac{1}{n}$ of a degree Centigrade is $\frac{1}{273\,n}$. The first factor $\frac{\delta\mu}{\mu-1}$, or $\frac{\text{difference of density}}{\text{density}}$, will therefore become $\frac{1}{H} - \frac{1}{273\,n}$, which, if we put $n = 540$ (corresponding to 1° Fahr. in 300 feet), and reckon H as 26,000, is approximately $\frac{1}{26000} - \frac{1}{147000}$ or $\frac{1}{H}\left(1 - \frac{1}{6}\right)$. The second factor of the expression for $\frac{1}{\rho}$ is unaffected. It appears, then, that decrease of temperature upwards at the rate of 1° C. in 540 feet, or 1° F. in 300 feet (which is the generally-received average), makes the curvature of the ray five-sixths of what it would be if the temperature were uniform.[1]

Combining this correction with correction II., it appears that, with a mean temperature of 10° C. or 50° F., and barometer at 760^{mm}, the curvature of a nearly horizontal ray (taking the earth's curvature as unity) is

$$\frac{1}{4\frac{1}{4}} \times \left(\frac{273}{283}\right)^2 \times \frac{5}{6} = \frac{1}{5{\cdot}5} \text{ nearly.}$$

This is in perfect agreement with observation, the received average (obtained as an empirical deduction from observation) being $\frac{1}{5}$ or $\frac{1}{6}$.

1087. Curvature of Inclined Rays.—Thus far we have been treating of nearly horizontal rays. To adapt our formula for $\frac{1}{\rho}$ ((2) § 1086) to the case of an oblique ray, we have merely to multiply it by $\cos\theta$,

[1] If the temperature decreases upwards at the rate of 1° C. in n feet, or 1° F. in n' feet, the first factor of the expression for $\frac{1}{\rho}$ (which would be $\frac{1}{H}$ at uniform temperature) becomes approximately $\frac{1}{H}\left(1 - \frac{96}{n}\right)$ or $\frac{1}{H}\left(1 - \frac{53}{n'}\right)$.

θ denoting the inclination of the ray to the horizontal, or the inclination of the wave-front to the vertical. For, if we still compare two points a foot apart, on the same wave-front, and in the same vertical plane with each other and with the ray, their difference of height will be the product of 1 foot by cos θ, and $\frac{\delta v}{v}$ will therefore be less than before in the ratio cos θ.

Hence it can be shown that the earth's curvature, so far from being the cause of terrestrial refraction, rather tends in ordinary circumstances to diminish it, by increasing the average obliquity of a ray joining two points at the same level.

The general formula for the curvature of a ray (lying in a vertical plane) at any point in its length, may be written

$$\begin{aligned}\frac{1}{\rho}&=\frac{1}{\mathrm{H}}\left(1-\frac{96}{n}\right)(\mu-1)\cos\theta\\&=\frac{1}{\mathrm{H}}\left(1-\frac{53}{n'}\right)(\mu-1)\cos\theta,\end{aligned}\tag{3}$$

n denoting the number of feet of ascent which give a decrease of 1° C., and n' the number of feet which give a decrease of 1° F. The unit of length for H and ρ may be anything we please.

1088. Astronomical Refraction.—Astronomical refraction, in virtue of which stars appear nearer the zenith than they really are, can be reduced to these principles; but it is simpler, in the case of stars not more than 70° or 80° from the zenith, to regard the earth and the layers of equal density in the atmosphere as plane, and to assume (§ 993) that the final result is the same as if the rays from the star were refracted at once out of vacuum into the horizontal stratum of air in which the observer's eye is situated. If z be the apparent and $z+h$ the true zenith distance, we shall thus have

$$\begin{aligned}\mu\sin z&=\sin(z+h)\\&=\sin z\cos h+\cos z\sin h\\&=\sin z+h\cos z,\text{ nearly,}\end{aligned}$$

whence

$$h=(\mu-1)\tan z.$$

1089. Mirage.—An appearance, as of water, is frequently seen in sandy deserts, where the soil is highly heated by the sun. The observer sees in the distance the reflection of the sky and of terrestrial objects, as in the surface of a calm lake. This phenomenon, which is called *mirage*, is explained by the heating and consequent rarefaction of the air in contact with the hot soil. The density,

within a certain distance of the ground, increases upwards, and rays traversing this portion of the air are bent upwards (Fig. 773), in accordance with the general rule that the concavity must be turned towards the denser side. Rays which were descending at a very slight inclination before entering this stratum of air may have their direction so much changed as to be bent up to an observer's eye, and the change of direction will be greatest for those rays which have

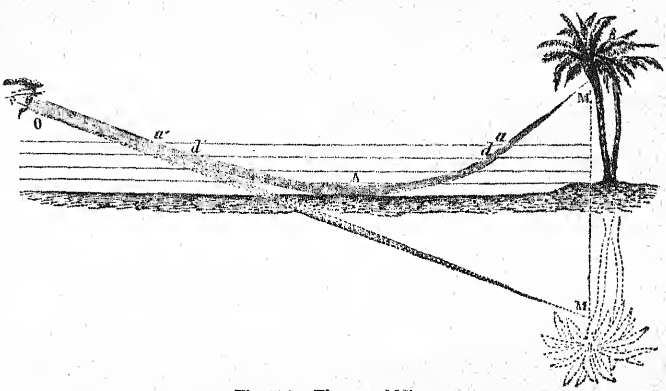


Fig. 773.—Theory of Mirage.

descended lowest; for these will not only have travelled for the greatest distance in the stratum, but will also have travelled through that part of it in which the change of density is most rapid. Hence, if we trace a pencil of rays from the observer's eye, we shall find that those of them which lie in the same vertical plane cross each other in traversing this stratum, and thus produce inverted images. If the stratum is thin in comparison with the height of the observer's eye, the appearance presented will be nearly equivalent to that produced by a mirror, while the objects thus reflected are also seen erect by higher rays which have not descended into the stratum where this action occurs.

A kind of inverted mirage is often seen across masses of calm water, and is called *looming;* images of distant objects, such as ships or hills, being seen in an inverted position immediately over the objects themselves. The explanation just given of the mirage of the desert will apply to this phenomenon also, if we suppose at a certain height, greater than that of the observer's eye, a layer of rapid transition from colder and denser air below to warmer and rarer air above.

An appearance similar to mirage may be obtained by gently depositing alcohol or methylated spirit upon water in a vessel with plate-glass sides. The spirit, though lighter, has a higher index of refraction than the water, and rays traversing the layer of transition are bent upwards. This layer accordingly behaves like a mirror when looked at very obliquely by an eye above it.[1]

1090. Curved Rays of Sound.—The reasoning of §§ 1084, 1086 can be applied, with a slight modification, to the propagation of sound.

Sound travels faster in warm than in cold air. On calm sunny afternoons, when the ground has become highly heated by the sun's rays, the temperature of the air is much higher near the ground than at moderate heights; hence sound bends upwards, and may thus become inaudible to observers at a distance by passing over their heads. On the other hand, on clear calm nights the ground is cooled by radiation to the sky, and the layers of air near the ground are colder than those above them; hence sound bends downwards, and may thus, by arching over intervening obstacles, become audible at distant points, which it could not reach by rectilinear propagation. This influence of temperature, which was first pointed out by Professor Osborne Reynolds, is one reason why sound from distant sources is better heard by night than by day.

A similar effect of wind had been previously pointed out by Professor Stokes. It is well known that sound is better heard with the wind than against it. This difference is due to the circumstance that wind is checked by friction against the earth, and therefore increases in velocity upwards. Sound travelling with the wind, therefore, travels fastest above, and sound travelling against the wind travels fastest below, its actual velocity being in the former case the sum, and in the latter the difference, of its velocity in still air and the velocity of the wind. The velocity of the wind is so much less than that of sound, that if uniform at all heights its influence on audibility would scarcely be appreciable.

1091. Calculation.—To calculate the curvature of a ray of sound due to variation of temperature with height, we may employ, as in § 1086, the formula $\frac{1}{\rho} = \frac{\delta v}{v}$, where δv denotes the difference of velocity for a difference of 1 foot in height. The value of v varies as $\sqrt{(1 + a\,t)}$, or approximately as $1 + \frac{1}{2}\,a\,t$, t denoting temperature, and a the co-

[1] A more complete discussion of the optics of mirage will be found in two papers by the editor of this work in the *Philosophical Magazine* for March and April, 1873, and in *Nature* for Nov. 19 and 26, 1874.

efficient of expansion, which is $\frac{1}{273}$. Hence if the velocity at 0° be denoted by 1, the value at $t°$ will be denoted by $1 + \frac{1}{2}\, a\, t$; and if the temperature varies by $\frac{1}{n}$ of a degree per foot, the value of $\frac{\delta v}{v}$ at temperatures near zero will be $\frac{a}{2n}$, that is, $\frac{1}{546n}$, and the radius of curvature will be $546\, n$ feet. This calculation shows that the bending is much more considerable for rays of sound than for rays of light.

1092. Diffraction Fringes.—When a beam of direct sunlight is admitted into a dark room through a narrow slit, a screen placed at any distance to receive it will show a line of white light, bordered with coloured fringes which become wider as the slit is narrowed. They also increase in width as the screen is removed further off. If they are viewed through a piece of red glass which allows only red rays to pass, they will appear as a succession of bands alternately bright and dark.

To explain their origin, we shall suppose the sun's rays (which may be reflected from an external mirror) to be perpendicular to the plane of the slit,[1] so that the wave-fronts are parallel to this plane, and we shall, in the first instance, confine our attention to light of a particular wave-length; for example, that of the light transmitted by the red glass. Then, if the slit be uniform through its whole length, the positions of the bright and dark bands will be governed by the following laws:—

1. The darkest parts will be at points whose distances from the two edges of the slit differ by an exact number of wave-lengths. If the difference be one wave-length, the light which arrives at any instant from different parts of the width of the slit is in all possible phases, and the resultant of the whole is zero. In fact, the disturbance produced by the nearer half of the slit cancels that produced by the remoter half. If the difference be n wave-lengths, we can divide the slit into n parts, such that the effect due to each part is thus *nil*.

2. The brightest parts will be at points whose distances from the two edges of the slit differ by an exact number of wave-lengths *plus*

[1] That is, to the plane of the two knife-edges by which the slit is bounded. This condition can only be strictly fulfilled for a single point on the sun's disc. Every point on the sun's surface sends out its own waves as an independent source; and waves from one point cannot interfere with waves from another. In the experiment as described in the text the fringes due to different parts of the sun's surface are all produced at once on the screen, and overlap each other.

a half. Let the difference be $n+\frac{1}{2}$; then we can divide the slit into n inefficient parts and one efficient part, this latter having only half the width of one of the others.[1]

Each colour of light has its own alternate bands of brightness and darkness, the distance from band to band being greatest for red and least for violet. The superposition of all the bands constitutes the coloured fringes which are seen.

This experiment furnishes the simplest answer to the objection formerly raised to the undulatory theory, that light is not able, like sound, to pass round an obstacle, but can only travel in straight lines. In this experiment light does pass round an obstacle, and turns more and more away from a straight line as the slit is narrowed.

When the slit is not exceedingly narrow, the light sent in oblique directions is quite insensible in comparison with the direct light, and no fringes are visible. "We have reason to think that when *sound* passes through a very large aperture, or when it is reflected from a large surface (which amounts nearly to the same thing), it is hardly sensible except in front of the opening, or in the direction of reflection."[2]

There are several other modes of producing diffraction fringes, which our limits do not permit us to notice. We proceed to describe the mode of obtaining a *pure spectrum* by diffraction. √

1093. Diffraction by a Grating.—If a piece of glass is ruled with parallel equidistant scratches (by means of a dividing engine and diamond point) at the rate of some hundreds or thousands to the inch, we shall find, on looking through it at a slit or other bright line (the glass being held so that the scratches are parallel to the slit), that a number of spectra are presented to view, ranged at nearly equal distances, on both sides of the slit. If the experiment is made under favourable circumstances, the spectra will be so pure as to show a number of Fraunhofer's lines.

Instead of viewing the spectra with the naked eye, we may with advantage employ a telescope, focussed on the plane of the slit; or we may project the spectra on a screen, by first placing a convex

[1] Each element of the length of the slit tends to produce a system of circular rings (the screen being supposed parallel to the plane of the slit). If the width of the slit is uniform, these systems will be precisely alike, and will have for their resultant a system of straight bands, parallel to the slit and touching the rings. These are the bands described in the text. Hence, to determine the illumination of any point of the screen, it is only necessary to attend, as in the text, to the nearest points of the two edges of the slit.

[2] Airy, *Undulatory Theory*. Art. 28.

lens so as to form an image of the slit (which must be very strongly illuminated) on the screen, and then interposing the ruled glass in the path of the beam.

A piece of glass thus ruled is called a *grating*.[1] A grating for diffraction experiments consists essentially of a number of parallel strips alternately transparent and opaque.

The distance between the "fixed lines" of the spectra, and the distance from one spectrum to the next, are found to depend on the distance of the strips measured from centre to centre, in other words, on the number of scratches to the inch, but not at all on the relative breadths of the transparent and opaque strips. This latter circumstance only affects the brightness of the spectra.

Diffraction spectra are of great practical importance—

1. As furnishing a uniform standard of reference in the comparison of spectra.

2. As affording the most accurate method of determining the wave-lengths of the different elementary rays of light.

1094. Principle of Diffraction Spectrum.—Let GG (Fig. 774) be a grating, receiving light from an infinitely[2] distant point lying in a direction perpendicular to the plane of the grating, so that the wave-fronts of the incident light are parallel to this plane. Let a convex lens L be placed on the other side of the grating, and let its axis make an acute angle θ with the rays incident on the grating. Then the light collected at its principal focus F consists of all the light incident upon the lens parallel to its axis. Let s denote the distance

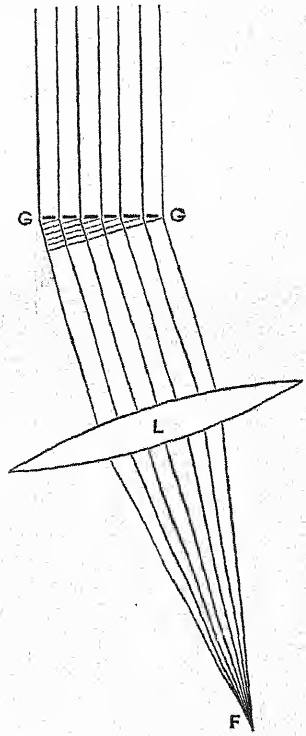


Fig. 774.
Principle of Diffraction Spectrum.

[1] Engraved glass gratings of sufficient size for spectroscopic purposes (say an inch square) are extremely expensive and difficult to procure. Lord Rayleigh has made numerous photographic copies of such gratings, and the copies appear to be equally effective with the originals.

[2] It is not *necessary* that the source should be infinitely distant (or the incident rays parallel); but this is the simplest case, and the most usual case in practice.

between the rulings, measured from centre to centre, so that if, for example, there are 1000 lines to the inch, s will be $\frac{1}{1000}$ of an inch; and suppose first that $s \sin\theta$ is exactly equal to the wave-length λ of one of the elementary kinds of light. Then, of all the light which falls upon the lens parallel to its axis, the left-hand portion in the figure is most retarded (having travelled farthest), and the right-hand portion least, the retardation, in comparing each transparent interval with the next, being constant, and equal to $s \sin\theta$, as is evident from an inspection of the figure. Now, for the particular kind of light for which $\lambda = s \sin\theta$, this retardation is exactly a wave-length, and all the transparent intervals send light of the same phase to the focus F; so that, if there are 1000 such intervals, the resultant amplitude of vibration at F is 1000 times the amplitude due to one interval alone. For light of any other wave-length this coincidence of phase will not exist. For example, if the difference between λ and $s \sin\theta$ is $\frac{1}{1000}\lambda$, the difference of phase between the lights received from the 1st and 2d intervals will be $\frac{1}{1000}\lambda$, between the 1st and 3d $\frac{2}{1000}\lambda$, between the 1st and 501st $\frac{500}{1000}\lambda$, or just half a wave-length, and so on. The 1st and 501st are thus in complete discordance, as are also the 2d and 502d, &c. Light of every wave-length except one is thus almost completely destroyed by interference, and the light collected at F consists almost entirely of the particular kind defined by the condition

$$\lambda = s \sin\theta. \qquad (1)$$

The purity of the diffraction spectrum is thus explained.

If a screen be held at F, with its plane perpendicular to the principal axis, any point on this screen a little to one side of F will receive light of another definite wave-length, corresponding to another direction of incidence on the lens, and a pure spectrum will thus be depicted on the screen.

1095. Practical Application.—In the arrangement actually employed for accurate observation, the lens L L is the object-glass of a telescope with a cross of spider-lines at its principal focus F. The telescope is first pointed directly towards the source of light, and is then turned to one side through a measured angle θ. Any fixed line of the spectrum can thus be brought into apparent coincidence with the cross of spider-lines, and its wave-length can be computed by the formula (1).

The spectrum to which formula (1) relates is called the *spectrum of the first order.*

There is also a spectrum of the second order, corresponding to values of θ nearly twice as great, and for which the equation is

$$2\lambda = s \sin \theta. \qquad (2)$$

For the spectrum of the third order, the equation is

$$3\lambda = s \sin \theta; \qquad (3)$$

and so on, the explanation of their formation being almost precisely the same as that above given. There are two spectra of each order, one to the right, and the other at the same distance to the left of the direction of the source. In Ångstrom's observations,[1] which are the best yet taken, all the spectra, up to the sixth inclusive, were observed, and numerous independent determinations of wave-length were thus obtained for several hundred of the dark lines of the solar spectrum.

The source of light was the infinitely distant image of an illuminated slit, the slit being placed at the principal focus of a collimator, and illuminated by a beam of the sun's rays reflected from a mirror.

The purity of a diffraction spectrum increases with the number of lines on the grating which come into play, provided that they are exactly equidistant; and may therefore be increased either by increasing the size of the grating, or by ruling its lines closer together. The gratings employed by Ångstrom were about $\frac{3}{4}$ of an inch square, the closest ruled having about 4500 lines, and the widest 1500.

As regards brightness, diffraction spectra are far inferior to those obtained by prisms. To give a maximum of light, the opaque intervals should be perfectly opaque, and the transparent intervals perfectly transparent; but even under the most favourable conditions, the whole light of any one of the spectra cannot exceed about $\frac{1}{10}$ of the light which would be received by directing the telescope to the slit. The greatest attainable intrinsic brightness in any part of a diffraction spectrum is thus not more than $\frac{1}{10}$ of the intrinsic brightness in the same part of a prismatic spectrum, obtained with the same slit, collimator, and observing telescope, and with the same angular separation of fixed lines. The brightness of the spectra partly depends upon the ratio of the breadths of the transparent and opaque intervals. In the case of the spectra of the first order, the best ratio is that of equality, and equal departures from equality in opposite directions give identical results; for example, if the breadth

[1] Ångstrom, *Recherches sur la Spectre solaire.* Upsal, 1868.

of the transparent intervals is to the breadth of the opaque either as 1 : 5 or as 5 : 1, it can be shown that the quantity of light in the first spectrum is just a quarter of what it would be with the breadths equal.

When a diffraction spectrum is seen with the naked eye, the cornea and crystalline of the eye take the place of the lens L L, and form a real image on the retina at F.

1096. Retardation Gratings.—If, instead of supposing the bars of the grating to be opaque, we suppose them to be transparent, but to produce a definite change of phase either by acceleration or retardation, the spectra produced will be the same as in the case above discussed, except as regards brightness. We may regard the effect as consisting of the superposition of two exactly coincident sets of spectra, one due to the spaces and the other to the bars. Any one of the resultant spectra may be either brighter or less bright than either of its components, according to the difference of phase between them. If the bars and spaces are equally transparent, the two superimposed spectra will be equally bright, and their resultant at any part may have any brightness intermediate between zero and four times that of either component.

1097. Reflection Gratings.—Diffraction spectra can also be obtained by reflection from a surface of speculum metal finely ruled with parallel and equidistant scratches. The appearance presented is the same as if the geometrical image of the slit (with respect to the grating regarded as a plane mirror) were viewed through the grating regarded as transparent.

1098. Standard Spectrum.—The simplicity of the law connecting wave-length with position, in the spectra obtained by diffraction, offers a remarkable contrast to the "irrationality" of the dispersion produced by prisms. Diffraction spectra may thus be fairly regarded as natural standards of comparison; and, in particular, the limiting form (if we may so call it) to which the diffraction spectra tend, as sin θ becomes small enough to be identified with θ, so that deviation becomes simply proportional to wave-length, is generally and deservedly accepted by spectroscopists as the *absolute standard of reference.* This limiting form is often briefly designated as "the diffraction spectrum;" it differs in fact to a scarcely appreciable extent from the first, or even the second and third spectra furnished in ordinary cases by a grating.

The diffraction spectrum differs notably from prismatic spectra in

the much greater relative extension of the red end. Owing to this circumstance, the brightest part of the diffraction spectrum of solar light is nearly in its centre.

The first three columns of numbers in the subjoined table indicate the approximate distances between the fixed lines B, D, E, F, G in certain prismatic spectra, and in the standard diffraction spectrum, the distance from B to G being in each case taken as 1000:—

	Flint-glass. Angle of 60°.	Bisulphide of Carbon. Angle of 60°.	Diffraction, or Difference of Wave-length.	Difference of Wave-frequency.
B to D, . .	220	194	381	278
D to E, . .	214	206	243	232
E to F, . .	192	190	160	184
F to G, . .	374	410	216	306
	1000	1000	1000	1000

In the standard diffraction spectrum, deviation is simply proportional to wave-length, and therefore the distance between two colours represents the difference of their wave-lengths. It has been suggested that a more convenient reference-spectrum would be constructed by assigning to each colour a deviation proportional to its wave-frequency (or to the reciprocal of its wave-length), so that the distance between two colours will represent the difference between their wave-frequencies. The result of thus disposing the fixed lines is shown in the last column of the above table. It differs from prismatic spectra in the same direction, but to a much less extent than the diffraction spectrum.

It has been suggested by Mr. Stoney as extremely probable, that the bright lines of spectra are in many cases harmonics of some one fundamental vibration. Three of the four bright lines of hydrogen have wave-frequencies exactly proportional to the numbers 20, 27, and 32; and in the spectrum of chloro-chromic acid all the lines whose positions have been observed (31 in number) have wave-frequencies which are multiples of one common fundamental.

47 **1099. Wave-lengths.**—Wave-lengths of light are commonly stated in terms of a unit of which 10^{10} make a metre,—hence called the *tenth-metre*. The following are the wave-lengths of some of the principal "fixed lines" as determined by Ångstrom:[1]—

[1] The wave-lengths of the spectral lines of all elementary substances will be found in Dr. W. M. Watts' *Index of Spectra;* and the wave-lengths and wave-frequencies of the dark lines in the solar spectrum, with the names of the substances to which many of them are due, will be found in the *British Association Report* for 1878 (Dublin), pp. 40–91.

WAVE-LENGTHS IN TENTH-METRES.

A	7604	E	5269
B	6867	F	4861
C	6562	G	4307
D_2	5895	H_1	3968
D_1	5889	H_2	3933

The velocity of light is 300 million metres per second, or 300×10^{16} tenth-metres per second. The number of waves per second for any colour is therefore 300×10^{16} divided by its wave-length as above expressed. Hence we find approximately:—

For A	395	millions of millions per second.
„ D	510	„ „ „
„ H	760	„ „ „

1100. Colours of Thin Films. Newton's Rings.—If two pieces of glass, with their surfaces clean, are brought into close contact, coloured fringes are seen surrounding the point where the contact is closest. They are best seen when light is obliquely reflected to the eye from the surfaces of the glass, and fringes of the complementary colours may be seen by transmitted light. A drop of oil placed on the surface of clean water spreads out into a thin film, which exhibits similar fringes of colour; and in general, a very thin film of any transparent substance, separating media whose indices of refraction are different from its own, exhibits colour, especially when viewed by obliquely reflected light. In the first experiment above-mentioned, the thin film is an air-film separating the pieces of glass. In soap-bubbles or films of soapy water stretched on rings, a similar effect is produced by a small thickness of water separating two portions of air.

The colours, in all these cases, when seen by reflected light, are produced by the mutual interference of the light reflected from the two surfaces of the thin film. An incident ray undergoes, as explained in § 992, a series of reflections and refractions; and we may thus distinguish, for light of any given refrangibility, several systems of waves, all of which originally came from the same source. These systems give by their interference a series of alternately bright and dark fringes; and when ordinary white light is employed, the fringes are broadest for the colours of greatest wave-length. Their superposition thus produces the observed colours. The colours seen by transmitted light may be similarly explained.

The first careful observations of these coloured fringes were made by Newton, and they are generally known as *Newton's rings.*

CHAPTER LXXV.

POLARIZATION AND DOUBLE REFRACTION.

1101. Polarization.—When a piece of the semi-transparent mineral called tourmaline is cut into slices by sections parallel to its axis, it is found that two of these slices, if laid one upon the other in a particular relative position, as A, B (Fig. 775), form an opaque combination. Let one of them, in fact, be turned round upon the other through various angles (Fig. 775). It will be found that the combination is most transparent in two positions differing by 180°, one of them $a\,b$ being the natural position which they originally occupied in the crystal; and that it is most opaque in the two positions at right angles to these. It is not necessary that the slices should be cut from the same crystal. Any two plates of tourmaline with their faces parallel to the axis of the crystals from which they were cut, will exhibit the same phenomenon. The experiment shows that light which has passed through one such plate is in a peculiar and so to speak unsymmetrical condition. It is said to be *plane-polarized*. According to the undulatory theory, a ray of common light contains vibrations in all planes passing through the ray, and a ray of plane-polarized light contains vibrations in one plane only. Polarized light cannot be distinguished from common light by the naked eye; and for all experiments in polarization two pieces of apparatus must be employed —one to produce polarization, and the other to show it. The former is called the *polarizer*, the latter the *analyser;* and every apparatus that serves for one of these purposes will also serve for the other. In the experiment above described, the plate next the eye is the

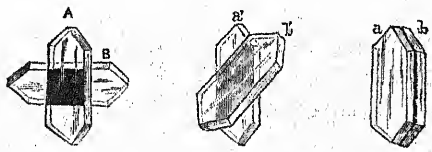


Fig. 775.—Tourmaline Plates.

analyser. The usual process in examining light with a view to test whether it is polarized, consists in looking at it through an analyser, and observing whether any change of brightness occurs as the analyser is rotated. When the light of the blue sky is thus examined, a difference of brightness can always be detected according to the position of the analyser, especially at the distance of about 90° from the sun. In all such cases there are two positions differing by 180°, which give a minimum of light, and the two positions intermediate between these give a maximum of light.

The extent of the changes thus observed is a measure of the completeness of the polarization of the light.

1102. Polarization by Reflection.—Transmission through tourmaline is only one of several ways in which light can be polarized. When a beam of light is reflected from a polished surface of glass, wood, ivory, leather, or any other non-metallic substance, at an angle of from 50° to 60° with the normal, it is more or less polarized, and in like manner a reflector composed of any of these substances may be employed as an analyser. In so using it, it should be rotated about an axis parallel to the incident rays which are to be tested, and the observation consists in noting whether this rotation produces changes in the amount of reflected light.

Malus' Polariscope (Fig. 776) consists of two reflectors A, B, one serving as polarizer and the other as analyser, each consisting of a pile of glass plates. Each of these reflectors can be turned about a horizontal axis; and the upper one (which is the analyser) can also be turned about a vertical axis, the amount of rotation being measured on the horizontal circle C C. To obtain the most powerful effects, each of the reflectors should be set at an angle of about 33° to the vertical, and a strong beam of common light should be allowed to fall upon the lower pile in such a direction as to be reflected vertically upwards. It will thus fall upon the centre of the upper pile, and the angles of incidence and reflection on both the piles will be about 57°. The observer looking into the upper pile, in such a

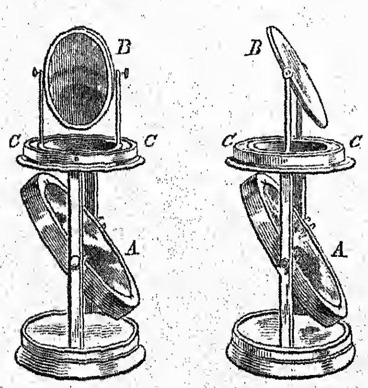


Fig. 776.—Malus' Polariscope.

direction as to receive the reflected beam, will find that, as the upper pile is rotated about a vertical axis, there are two positions (differing by 180°) in which he sees a black spot in the centre of the field of view, these being the positions in which the upper pile refuses to reflect the light reflected to it from the lower pile. They are 90° on either side of the position in which the two piles are parallel; this latter, and the position differing from it by 180°, being those which give a maximum of reflected light.

For every reflecting substance there is a particular angle of incidence which gives a maximum of polarization in the reflected light. It is called the *polarizing angle* for the substance, and its tangent is always equal to the index of refraction of the substance; or what amounts to the same thing, it is that particular angle of incidence which is the complement of the angle of refraction, so that the refracted and reflected rays are at right angles.[1] This important law was discovered experimentally by Sir David Brewster.

The reflected ray under these circumstances is in a state of almost complete polarization; and the advantage of employing a *pile* of plates consists merely in the greater intensity of the reflected light thus furnished. The transmitted light is also polarized; it diminishes in intensity, but becomes more completely polarized, as the number of plates is increased. The reflected and the transmitted light are in fact mutually complementary, being the two parts into which common light has been decomposed; and their polarizations are accordingly opposite, so that, if both the transmitted and reflected beams are examined by a tourmaline, the maxima of obscuration will be obtained by placing the axis of the tourmaline in the one case parallel and in the other perpendicular to the plane of incidence.

It is to be noted that what is lost in reflection is gained in transmission, and that polarization never favours reflection at the expense of transmission.

1103. Plane of Polarization.—That particular plane in which a ray of polarized light, incident at the polarizing angle, is most copiously reflected, is called the *plane of polarization* of the ray. When the polarization is produced by reflection, the plane of reflection is the

[1] Adopting the indices of refraction given in the table § 986, we find the following values for the polarizing angle for the undermentioned substances:—

Diamond, . . .	67° 43′ to 70° 3′	Crown-glass, . .	56° 51′ to 57° 23′
Flint-glass, . .	57° 36′ to 58° 40′	Pure Water, ...	53° 11′

Polarized light - useful for discovering adulteration in sugar

plane of polarization. According to Fresnel's theory, which is that generally received, the vibrations of light polarized in any plane are perpendicular to that plane (§ 1115). The vibrations of a ray reflected at the polarizing angle are accordingly to be regarded as perpendicular to the plane of incidence and reflection, and therefore as parallel to the reflecting surface.

1104. Polarization by Double Refraction.—We have described in § 998 some of the principal phenomena of double refraction in uniaxal crystals. We have now to mention the important fact that the two rays furnished by double refraction are polarized, the polarization in this case being more complete than in any of the cases thus far discussed. On looking at the two images through a plate of tourmaline, or any other analyser, it will be found that they undergo great variations of brightness as the analyser is rotated, one of them becoming fainter whenever the other becomes brighter, and the maximum brightness of either being simultaneous with the absolute extinction of the other. If a second piece of Iceland-spar be used as the analyser, four images will be seen, of which one pair become dimmer as the other pair become brighter, and either of these pairs can be extinguished by giving the analyser a proper position.

1105. Theory of Double Refraction.—The existence of double refraction admits of a very natural explanation on the undulatory theory. In uniaxal crystals it is assumed that the elasticity of the luminiferous æther is the same for all vibrations executed in directions perpendicular to the axis; and that, for vibrations in other directions, the elasticity varies solely according to the inclination of the direction of vibration to the axis. There are two classes of doubly-refracting uniaxal crystals, called respectively *positive* and *negative*. In the former the elasticity for vibrations perpendicular to the axis is a maximum; in the latter it is a minimum. Iceland-spar belongs to the latter class; and as small elasticity implies slow propagation, a ray propagated by vibrations perpendicular to the axis will, in this crystal, travel with minimum velocity; while the most rapid propagation will be attained by rays whose vibrations are parallel to the axis.

Consider any plane oblique to the axis. Through any point in this plane we can draw one line perpendicular to the axis; and the line at right angles to this will have smaller inclination to the axis than any other line in the plane. These two lines are the directions of least and greatest resistance to vibration; the former is the direc-

tion of vibration for an ordinary, and the latter for an extraordinary ray. The velocity of propagation is the same for the ordinary rays in all directions in the crystal, so that the wave-surface for these is spherical; but the velocity of propagation for the extraordinary rays differs according to their inclination to the axis, and their wave-surface is a spheroid whose polar diameter is equal to the diameter of the aforesaid sphere. The sphere and spheroid touch one another at the extremities of this diameter (which is parallel to the axis of the crystal), and the ordinary and extraordinary rays coincide both in direction and velocity along this common diameter. The general construction for the path of the extraordinary ray is due to Huygens, and has been described in § 1081, Fig. 770, where C A is the incident and A F the refracted ray.

When the plane of incidence contains the axis, the spheroid will be symmetrical with respect to this plane; and, therefore, when we draw the tangent plane E F perpendicular to the plane of incidence (as directed in the construction) the point of contact F will lie in the plane of incidence.

Another special case is that in which the plane of incidence is perpendicular to the axis, and the refracting surface parallel to the axis. In this case also the spheroid will be symmetrical with respect to the plane of incidence, which will in fact be the equatorial plane of the spheroid, and the point of contact F will as before lie in this plane. Moreover since the section is equatorial it is a circle, and hence, as shown in Fig. 771, the law of sines will be applicable. The ratio of the sines of the angles of incidence and refraction for this particular case is called the *extraordinary index* of refraction for the crystal. It is the ratio of the velocity in air to the velocity along an equatorial radius of the spheroid.

In general, the spheroid is not symmetrical with respect to the plane of incidence, and the refracted ray A F does not lie in this plane.

Tourmaline, like Iceland-spar, is a negative uniaxal crystal; and its use as a polarizer depends on the property which it possesses of absorbing the ordinary much more rapidly than the extraordinary ray, so that a thickness which is tolerably transparent to the latter is almost completely opaque to the former.

1106. Nicol's Prism.—One of the most convenient and effective contrivances for polarizing light, or analysing it when polarized, is that known, from the name of its inventor, as Nicol's prism. It is

made by slitting a rhomb of Iceland-spar along a diagonal plane *a c b d* (Fig. 777), and cementing the two pieces together in their natural position by Canada balsam, a substance whose refractive index is intermediate between the ordinary and extraordinary indices of the crystal.[1] A ray of common light S I undergoes double refraction on entering the prism. Of the two rays thus formed, the ordinary ray is totally reflected on meeting the first surface of the balsam, and passes out at one side of the crystal, as *o* O; while the extraordinary ray is transmitted through the balsam as through a parallel plate, and finally emerges at the end of the prism, in the direction *e* E, parallel to the original direction S I. This apparatus has nearly all the convenience of a tourmaline plate, with the advantages of much greater transparency and of complete polarization.

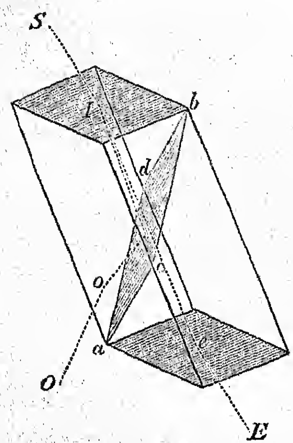


Fig. 777.—Nicol's Prism.

In Foucault's prism, which is extensively used instead of Nicol's, the Canada balsam is omitted, and there is nothing but air between the two pieces. This change has the advantage of shortening the prism (because the critical angle of total reflection depends on the relative index of refraction of the two media), but gives a smaller field of view, and rather more loss of light by reflection.

1107. Colours produced by Elliptic Polarization.—Very beautiful colours may be produced by the peculiar action of polarized light. For example, if a piece of selenite (crystallized gypsum) about the thickness of paper, is introduced between the polarizer and analyser of any polarizing arrangement, and turned about into different directions, it will in some positions appear brightly coloured, the colour being most decided when the analyser is in either of the two critical positions which give respectively the greatest light and the greatest darkness. The colour is changed to its complementary by

[1] *a* and *b* are the corners at which three equal obtuse angles meet (§ 999). The ends of the rhomb which are shaded in the figure are rhombuses. Their diagonals drawn through *a* and *b* respectively will lie in one plane, which will contain the axis of the crystal, and will cut the plane of section *a c b d* at right angles. The length of the rhomb is about three and a half times its breadth.

rotating the analyser through a right angle; but rotation of the piece of selenite, when the analyser is in either of the critical positions, merely alters the depth of the colour without changing its tint, and in certain critical positions of the selenite there is a complete absence of colour. Thicker plates of selenite restore the light when extinguished by the analyser, but do not show colour.

1108. Explanation.—The following is the explanation of these appearances. Let the analyser be turned into such a position as to produce complete extinction of the plane-polarized light which comes to it from the polarizer; and let the plane of polarization and the plane perpendicular thereto (and parallel to the polarized rays) be called the two *planes of reference.* Let the slice of selenite be laid so that the polarized rays pass through it normally. Then there are two directions, at right angles to each other, which are the directions of greatest and least elasticity in the plane of the slice. Unless the slice is laid so that these directions coincide with the two planes of reference, the plane-polarized light which is incident upon it will be broken up into two rays, one of which will traverse it more rapidly than the other. Referring to the diagram of Lissajous' figures (Fig. 634), let the sides of the rectangle be the directions of greatest and least elasticity, and let the diagonal line in the first figure be the direction of the vibrations of an incident ray,—this diagonal accordingly lies in one of the two planes of reference. In traversing the slice, the component vibrations in the directions of greatest and least elasticity will be propagated with unequal velocities; and if the incident ray be homogeneous, the emergent light will be elliptically polarized; that is to say, its vibrations, instead of being rectilinear, will be elliptic, precisely on the principle[1] of Blackburn's pendulum (§ 924). The shape of the ellipse depends, as in the case of Lissajous' figures, on the amount of retardation of one of the two component vibrations as compared with the other, and this is directly proportional to the thickness of the slice. The analyser resolves these elliptic vibrations into two rectilinear components parallel and perpendicular to the original direction of vibration, and suppresses one of these components, so that only the other remains.

[1] The principle is that, whereas displacement of a particle parallel to either of the sides of the rectangle calls out a restoring force directly opposite to the displacement, displacement in any other direction calls out a restoring force inclined to the direction of displacement, being in fact the resultant of the two restoring forces which its two components parallel to the sides of the rectangle would call out.

Thus if the ellipse in the annexed figure (Fig. 778) represent the vibrations of the light as it emerges from the selenite, and CD, E F be tangents parallel to the original direction of vibration, the perpendicular distance between these tangents, A B, is the component vibration which is not suppressed when the analyser is so turned that all the light would be suppressed if the selenite were removed. By rotating the analyser, we shall obtain vibrations of various amplitudes, corresponding to the distances between parallel tangents drawn in various directions.

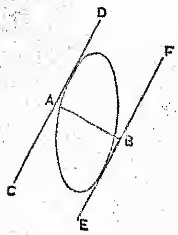


Fig. 778.—Colours of Selenite Plates.

For a certain thickness of selenite the ellipse may become a circle, and we have thus what is called *circularly polarized* light, which is characterized by the property that rotation of the analyser produces no change of intensity. Circularly polarized light is not however identical with ordinary light; for the interposition of an additional thickness of selenite converts it into elliptically (or in a particular case into plane) polarized light (§ 1114).

The above explanation applies to homogeneous light. When the incident light is of various refrangibilities, the retardation of one component upon the other is greatest for the rays of shortest wave-length. The ellipses are accordingly different for the different elementary colours, and the analyser in any given position will produce unequal suppression of different colours. But since the component which is suppressed in any one position of the analyser, is the component which is not suppressed when the analyser is turned through a right angle, the light yielded in the former case *plus* the light yielded in the latter must be equal to the whole light which was incident on the selenite.[1] Hence the colours exhibited in these two positions must be complementary.

It is necessary for the exhibition of colour in these experiments that the plate of selenite should be very thin, otherwise the retardation of one component vibration as compared with the other will be greater by several complete periods for violet than for red, so that the ellipses will be identical for several different colours, and the total non-suppressed light will be sensibly white in all positions of the analyser.

[1] We here neglect the light absorbed and scattered; but the loss of this does not sensibly affect the *colour* of the whole. It is to be borne in mind that the intensity of light is measured by the *square* of the amplitude, and is therefore the simple sum of the intensities of its two components when the resolution is rectangular.

Two thick plates may however be so combined as to produce the effect of one thin plate. For example, two selenite plates, of nearly equal thickness, may be laid one upon the other, so that the direction of greatest elasticity in the one shall be parallel to that of least elasticity in the other. The resultant effect in this case will be that due to the difference of their thicknesses. Two plates so laid are said to be *crossed*.

1109. Colours of Plates perpendicular to Axis.—A different class of

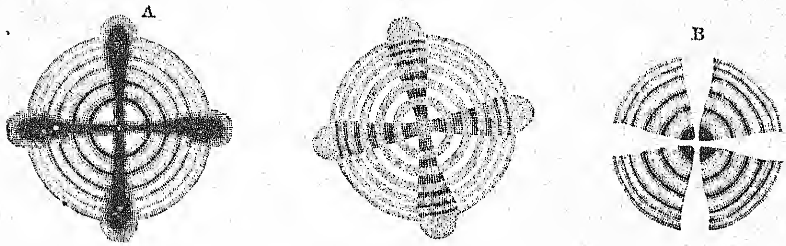


Fig. 779.—Rings and Cross.

appearances are presented when a plate, cut from a uniaxal crystal by sections perpendicular to the axis, is inserted between the polarizer and the analyser. Instead of a broad sheet of uniform colour, we have now a system of coloured rings, interrupted, when the analyser is in one of the two critical positions, by a black or white cross, as at A, B (Fig. 779).

1110. Explanation.—The following is the explanation of these appearances. Suppose, for simplicity, that the analyser is a plate of tourmaline held close to the eye. Then the light which comes to the eye from the nearest point of the plate under examination (the foot of a perpendicular dropped upon it from the eye), has traversed the plate normally, and therefore parallel to its optic axis. It has therefore not been resolved into an ordinary and an extraordinary ray, but has emerged from the plate in the same condition in which it entered, and is therefore black, gray, or white according to the position of the analyser, just as it would be if the plate were removed. But the light which comes obliquely to the eye from any other part of the plate, has traversed the plate obliquely, and has undergone double refraction. Let E (Fig. 780) be the position

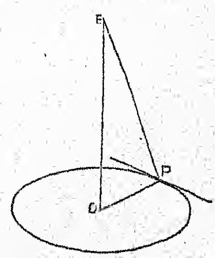


Fig. 780.
Theory of Rings and Cross.

of the eye, E O a perpendicular on the plate, P a point on the circumference of a circle described about O as centre. Then, since E O is parallel to the axis of the plate, the direction of vibration for the ordinary ray at P is perpendicular to the plane E O P, and is tangential to the circle. The direction of vibration for the extraordinary ray lies in the plane E O P, is nearly perpendicular to E O (or to the axis), if the angle O E P is small, and deviates more from perpendicularity to the axis as the angle O E P increases. Both for this reason, and also on account of the greater thickness traversed, the retardation of one ray upon the other is greater as P is taken further from O; and from the symmetry of the circumstances, it must be the same at the same distance from O all round. In consequence of this retardation, the light which emerges at P in the direction PE is elliptically polarized; and by the agency of the analyser it is accordingly resolved into two components, one of which is suppressed. With homogeneous light, rings alternately dark and bright would thus be formed at distances from O corresponding to retardations of $0, \frac{1}{2}, 1, 1\frac{1}{2}, 2, 2\frac{1}{2}, \ldots$ complete periods; and it can be shown that the radii of these rings would be proportional to the numbers $0, \sqrt{1}, \sqrt{2}, \sqrt{3}, \sqrt{4}, \sqrt{5}, \sqrt{6}: \ldots$ The rings are larger for light of long than of short wave-length; and the coloured rings actually exhibited when white light is employed, are produced by the superposition of all the systems of monochromatic rings. The monochromatic rings for red light are easily seen by looking at the actual rings through a piece of red glass.

Let O, P, Fig. 781, be the same points which were denoted by these letters in Fig. 780, and let A B be the direction of vibration of the light incident on the crystal at P. Draw A C, D B parallel to O P, and complete the rectangle A C B D. Then the length and breadth of this rectangle are approximately the directions of vibration of the two components, one of which loses upon the other in traversing the crystal. The vibration of the emergent ray is represented by an ellipse inscribed in the rectangle A C B D (§ 922, note 2); and when the loss is half a period, this ellipse shrinks into a straight line, namely, the diagonal C D. Through C and D draw lines parallel to A B; then the distance between these parallels

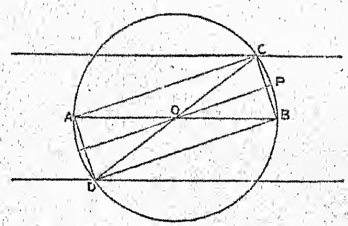


Fig. 781.—Theory of Rings and Cross.

represents the double amplitude of the vibration which is transmitted when there has been a retardation of half a period, and is greater than the distance between the tangents in the same direction to any of the inscribed ellipses. A retardation of another half period will again reduce the inscribed ellipse to the straight line A B, as at first. The position D C corresponds to the brightest and A B to the darkest part of any one of the series of rings for a given wave-length of light, the analyser being in the position for suppressing all the light if the crystal were removed. When the analyser is turned into the position at right angles to this, A B corresponds to the brightest, and D C to the darkest parts of the rings. It is to be remembered that amount of retardation depends upon distance from the centre of the rings, and is the same all round. The two diagonals of our rectangle therefore correspond to different sizes of rings.

If the analyser is in such a position with respect to the point P considered, that the suppressed vibration is parallel to one of the sides of the rectangle (in other words, if O P, or a line perpendicular to O P, is the direction of suppression) the retardation of one component upon the other has no influence, inasmuch as one of the two components is completely suppressed and the other is completely transmitted. There are, accordingly, in all positions of the analyser, a pair of diameters, coinciding with the directions of suppression and non-suppression, which are alike along their whole length and free from colour.

Again if P is situated at B or at 90° from B, the corner C of the rectangle coincides with B or with A, and the rectangle, with all its inscribed ellipses, shrinks into the straight line A B. The two diameters coincident with and perpendicular to A B are therefore alike along their whole length and uncoloured.

The two colourless crosses which we have thus accounted for, one of them turning with the analyser and the other remaining fixed with the polarizer, are easily observed when the analyser is not near the critical positions. In the critical positions, the two crosses come into coincidence; and these are also the positions of maximum blackness or maximum whiteness for the two crosses considered separately. Hence the conspicuous character of the cross in either of these positions, as represented at A, B, Fig. 779. As the analyser is turned away from these positions, the cross at first turns after it with half its angular velocity, but soon breaks up into rings, some-

what in the manner represented at C, which corresponds to a position not differing much from A.

1111. **Biaxal Crystals.** — Crystals may be divided optically into three classes:—

1. Those in which there is no distinction of different directions, as regards optical properties. Such crystals are said to be optically *isotropic*.

2. Those in which the optical properties are the same for all directions equally inclined to one particular direction called the optic axis, but vary according to this inclination. Such crystals are called *uniaxal*.

3. All remaining crystals (excluding compound and irregular formations) belong to the class called *biaxal*. In any homogeneous elastic solid, there are three cardinal directions called *axes of elasticity*, possessing the same distinctive properties which belong to the two principal planes of vibration in Blackburn's pendulum (§ 924); that is to say, if any small portion of the solid be distorted by forcibly displacing one of its particles in one of these cardinal directions, the forces of elasticity thus evoked tend to urge the particle *directly* back; whereas displacement in any other direction calls out forces whose resultant is generally oblique to the direction of displacement, so that when the particle is released it does not fly back through the position of equilibrium, but passes on one side of it, just as the bob of Blackburn's pendulum generally passes beside and not through the lowest point which it can reach.

In biaxal crystals, the resistances to displacement in the three cardinal directions are all unequal; and this is true not only for the crystalline substance itself, but also for the luminiferous æther which pervades it, and is influenced by it.[1] The construction given by Fresnel for the wave-surface in any crystal is as follows:—First take an ellipsoid, having its axis parallel to the three cardinal directions, and of lengths depending on the particular crystalline substance considered. Then let any plane sections (which will of course be ellipses) be made through the centre of this ellipsoid, let normals to them be drawn through the centre, and on each normal let points be taken at distances from the centre equal to the greatest and least radii of the corresponding section. The locus of these points is the complete wave-surface, which consists of two sheets cutting one

[1] The cardinal directions are however believed not to be the same for the æther as for the material of the crystal.

another at four points. These four points of intersection are situated upon the normals to the two *circular sections* of the ellipsoid, and the two *optic axes*, from which *biaxal* crystals derive their name, are closely related to these two circular sections. The optic axes are the directions of *single wave-velocity*, and the normals to the two circular sections are the directions of *single ray-velocity*. The direction of advance of a wave is always regarded as normal to the front of the wave, whereas the direction of a ray (defined by the condition of traversing two apertures placed in its path) always passes through the centre of the wave-surface, and is not in general normal to the front. Both these pairs of directions of single velocity are in the plane which contains the greatest and least axes of the ellipsoid.

When two axes of the ellipsoid are equal, it becomes a spheroid, and the crystal is uniaxal. When all three axes are equal, it becomes a sphere, and the crystal is isotropic.

Experiment has shown that biaxal crystals expand with heat unequally in three cardinal directions, so that in fact a spherical piece of such a crystal is changed into an ellipsoid[1] when its temperature is raised or lowered. A spherical piece of a uniaxal crystal in the same circumstances changes into a spheroid; and a spherical piece of an isotropic crystal remains a sphere.

It is generally possible to determine to which of the three classes a crystal belongs, from a mere inspection of its shape as it occurs in nature. Isotropic crystals are sometimes said to be symmetrical about a point, uniaxal crystals about a line, biaxal crystals about neither. The following statement is rather more precise:—

If there is one and only one line about which if the crystal be rotated through 90° or else through 120° the crystalline form remains in its original position, the crystal is uniaxal, having that line for the axis. If there is more than one such line, the crystal is isotropic, while, if there is no such line, it is biaxal. Even in the last case, if there exist a plane of crystalline symmetry, such that one half of the crystal is the reflected image of the other half with respect to this plane, it is also a plane of optical symmetry, and one of the three cardinal directions for the æther is perpendicular to it.[2]

[1] This fact furnishes the best possible definition of an ellipsoid for persons unacquainted with solid geometry.

[2] The optic axes either lie in the plane of symmetry, or lie in a perpendicular plane and are equally inclined to the plane of symmetry.

For the precise statement here given, the Editor is indebted to Professor Stokes.

Glass, when in a strained condition, ceases to be isotropic, and if inserted between a polarizer and an analyser, exhibits coloured streaks or spots, which afford an indication of the distribution of strain through its substance. The experiment is shown sometimes with unannealed glass, which is in a condition of permanent strain, sometimes with a piece of ordinary glass which can be subjected to force at pleasure by turning a screw. Any very small portion of a piece of strained glass has the optical properties of a crystal, but different portions have different properties, and hence the glass as a whole does not behave like one crystal.

The production of colour by interposition between a polarizer and an analyser, is by far the most delicate test of double refraction. Many organic bodies (for example, grains of starch) are thus found to be doubly refracting; and microscopists often avail themselves of this means of detecting diversities of structure in the objects which they examine.

1112. Rotation of Plane of Polarization.—When a plate of quartz (rock-crystal), even of considerable thickness, cut perpendicular to the axis, is interposed between the polarizer and analyser, colour is exhibited, the tints changing as the analyser is rotated; and similar effects of colour are produced by employing, instead of quartz, a solution of sugar, inclosed in a tube with plane glass ends.

If homogeneous light is employed, it is found that if the analyser is first adjusted to produce extinction of the polarized light, and the quartz or saccharine solution is then introduced, there is a partial restoration of light. On rotating the analyser through a certain angle, there is again complete extinction of the light; and on comparing different plates of quartz, it will be found that the angle through which the analyser must be rotated is proportional to the thickness of the plate. In the case of solutions of sugar, the angle is proportional jointly to the length of the tube and the strength of the solution.

The action thus exerted by quartz or sugar is called *rotation of the plane of polarization*, a name which precisely expresses the observed phenomena. In the case of ordinary quartz, and solutions of sugar-candy, it is necessary to rotate the analyser in the direction of watch-hands as seen by the observer, and the rotation of the plane of polarization is said to be *right-handed.* In the case of what is called *left-handed* quartz, and of solutions of non-crystallizable sugar, the rotation of the plane of polarization is in the

opposite direction, and the observer must rotate the analyser against watch-hands.

The amount of rotation is different for the different elementary colours, and has been found to be inversely as the square of the wave-length. Hence the production of colour.

1113. Magneto-optic Rotation.—Faraday made the remarkable discovery that the plane of polarization can be rotated in certain circumstances by the action of magnetism. Let a long rectangular piece of "heavy-glass" (silico-borate of lead) be placed longitudinally between the poles of the powerful electro-magnet represented in Fig. 445 (page 683), which is for this purpose made hollow in its axis, so that an observer can see through it from end to end. Let a Nicol's prism be fitted into one end of the magnet, to serve as polarizer, and another into the other end to serve as analyser, and let one of them be turned till the light is extinguished. Then, as long as no current is passed round the electro-magnet, the interposition of the heavy-glass will produce no effect; but the passing of a current while the heavy-glass is in its place between the poles, produces rotation of the plane of polarization in the same direction as that in which the current circulates. The amount of rotation is directly as the strength of current, and directly as the length of heavy-glass traversed by the light. Flint-glass gives about half the effect of heavy-glass, and all transparent solids and liquids exhibit an effect of the same kind in a more or less marked degree.

A steel magnet, if extremely powerful, may be used instead of an electro-magnet; and in all cases, to give the strongest effect, the lines of magnetic force should coincide with the direction of the transmitted ray.

Faraday regarded these phenomena as proving the direct action of magnetism upon light; but it is now more commonly believed that the direct effect of the magnetism is to put the particles of the transparent body in a peculiar state of strain, to which the observed optical effect is due.

In every case tried by Faraday, the direction of the rotation was the same as the direction in which the current circulated; but certain substances[1] have since been found which give rotation against the current. The law for the relative amounts of rotation of different colours is approximately the same as in the case of quartz.

[1] One such substance is a solution of Fe^2Cl^3 (old notation) in methylic (not methylated) alcohol.

The direction of rotation is with watch-hands as seen from one end of the arrangement, and against watch-hands as seen from the other; so that the same piece of glass, in the same circumstances, behaves like right-handed quartz to light entering it at one end, and like left-handed quartz to light entering it at the other.

The rotatory power of quartz and sugar appears to depend upon a certain unsymmetrical arrangement of their molecules, an arrangement somewhat analogous to the thread of a screw; right-handed and left-handed screws representing the two opposite rotatory powers. It is worthy of note that the two kinds of quartz crystallize in different forms, each of which is unsymmetrical, one being like the image of the other as seen in a looking-glass. Pasteur has conducted extremely interesting researches into the relations existing between substances which, while in other respects identical or nearly identical, differ as regards their power of producing rotation. For the results we must refer to treatises on chemistry.

Dr. Kerr has recently obtained rotation of the plane of polarization by reflection from intensely magnetized iron. In some of the experiments the direction of magnetization was normal, and in others parallel to the reflecting surface.

1114. Circular Polarization. Fresnel's Rhomb.—We have explained in § 1108 the process by which elliptic polarization is brought about, when plane-polarized light is transmitted through a thin plate of selenite. To obtain circular polarization (which is merely a case of elliptic), the plate must be of such thickness as to retard one component more than the other by a *quarter of a wave-length*, and must be laid so that the directions of the two component vibrations make angles of 45° with the plane of polarization. Plates specially prepared for this purpose are in general use, and are called *quarter-wave plates.* They are usually of mica, which differs but little in its properties from selenite. It is impossible, however, in this way to obtain complete circular polarization of ordinary white light, since different thicknesses are required for light of different wave-lengths, the thickness which is appropriate for violet being too small for red.

Fresnel discovered that plane-polarized light is elliptically polarized by *total internal reflection* in glass, whenever the plane of polarization of the incident light is inclined to the plane of incidence. The rectilinear vibrations of the incident light are in fact resolved into two components, one of them in, and the other per-

pendicular to, the plane of incidence; and one of these is retarded with respect to the other in the act of reflection, by an amount depending on the angle of incidence. He determined the magnitude of this angle for which the retardation is precisely $\frac{1}{8}$ of a wave-length; and constructed a *rhomb*, or oblique parallelepiped of glass (Fig. 782), in which a ray, entering normally at one end, undergoes two successive reflections at this angle (about 55°), the plane of reflection being the same in both. The total retardation of one component on the other is thus $\frac{1}{4}$ of a wave-length; and if the rhomb is in such a position that the plane in which the two reflections take place is at an angle of 45° to the plane of polarization of the incident light, the emergent light is circularly polarized. The effect does not vary much with the wave-length, and sensibly white circularly polarized light can accordingly be obtained by this method.

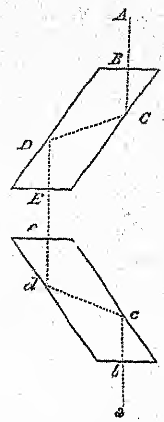


Fig. 782.
Two Fresnel's Rhombs.

When circularly polarized light is transmitted through a Fresnel's rhomb, or through a quarter-wave plate, it becomes plane-polarized, and we have thus a simple mode of distinguishing circularly polarized light from common light; for the latter does not become polarized when thus treated. Two quarter-wave plates, or two Fresnel's rhombs, may be combined either so as to assist or to oppose one another. By the former arrangement, which is represented in Fig. 782, we can convert plane-polarized light into light polarized in a perpendicular plane, the final result being therefore the same as if the plane of polarization had been rotated through 90°. The several steps of the process are illustrated by the five diagrams of Fig. 783,

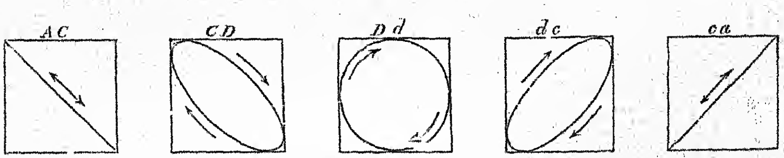


Fig. 783.—Form of Vibration in traversing the Rhombs.

which represent the vibrations of the five portions AC, CD, D*d*, *dc*, *ca* of the ray which traverses the two rhombs in the preceding figure. The sides of the square are parallel to the directions of resolution; the initial direction of vibration is one diagonal of the square, and

the final direction is the other diagonal; a gain or loss of half a complete vibration on the part of either component being just sufficient to effect this change.

1115. Direction of Vibration of Plane-polarized Light.—The plane of polarization of plane-polarized light may be defined as the plane in which it is most copiously reflected. It is perpendicular to the plane in which the light refuses to be reflected (at the polarizing angle); and is identical with the original plane of reflection, if the polarization was produced by reflection. This definition is somewhat arbitrary, but has been adopted by universal consent.

When light is polarized by the double refraction of Iceland-spar, or of any other uniaxal crystal, it is found that the plane of polarization of the ordinary ray is the plane which contains the axis of the crystal. But the distinctive properties of the ordinary ray are most naturally explained by supposing that its vibrations are perpendicular to the axis. Hence we conclude that the direction of vibration in plane-polarized light is normal to the so-called plane of polarization, and therefore that, in polarization by reflection, the vibrations of the reflected light are parallel to the reflecting surface.

This is Fresnel's doctrine. MacCullagh, however, reversed this hypothesis, and maintained that the direction of vibration is *in* the plane of polarization. Both theories have been ably expounded; but Stokes contrived a crucial experiment in diffraction, which confirmed Fresnel's view;[1] and in his classical paper on "Change of Refrangibility," he has deduced the same conclusion from a consideration of the phenomena of the polarization of light by reflection from excessively fine particles of solid matter in suspension in a liquid.[2]

1116. Vibrations of Ordinary Light.—Ordinary light agrees with circularly polarized light in always yielding two beams of equal intensity when subjected to double refraction; but it differs from circularly polarized light in not becoming plane-polarized by transmission through a Fresnel's rhomb or a quarter-wave plate. What, then, can be the form of vibration for common light? It is probably very irregular, consisting of ellipses of various sizes, positions, and forms (including circles and straight lines), rapidly succeeding one another. By this irregularity we can account for the fact that beams of light from different sources (even from different points of the same flame, or from different parts of the sun's disc), cannot, by

[1] *Cambridge Transactions.* 1850.
[2] *Philosophical Transactions,* 1852; pp. 530, 531.

any treatment whatever, be made to exhibit the phenomena of mutual interference; and for the additional fact that the two rectangular components into which a beam of common light is resolved by double refraction, cannot be made to interfere, even if their planes of polarization are brought into coincidence by one of the methods of rotation above described.

Certain phenomena of interference show that a few hundred consecutive vibrations of common light may be regarded as similar; but as the number of vibrations in a second is about 500 millions of millions, there is ample room for excessive diversity during the time that one impression remains upon the retina.

1117. Polarization of Radiant Heat.—The fundamental identity of radiant heat and light is confirmed by thermal experiments on polarization. Such experiments were first successfully performed by Forbes in 1834, shortly after Melloni's invention of the thermo-multiplier. He first proved the polarization of heat by tourmaline; next by transmission through a bundle of very thin mica plates, inclined to the transmitted rays; and afterwards by reflection from the multiplied surfaces of a pile of thin mica plates placed at the polarizing angle. He next succeeded in showing that polarized heat, even when quite obscure, is subject to the same modifications which doubly refracting crystallized bodies impress upon light, by suffering a beam of heat, after being polarized by transmission, to pass through an interposed plate of mica, serving the purpose of the plate of selenite in the experiment of § 1107, the heat traversing a second mica bundle before it was received on the thermo-pile. As the interposed plate was turned round in its own plane, the amount of heat shown by the galvanometer was found to fluctuate just as the amount of light received by the eye under similar circumstances would have done. He also succeeded in producing circular polarization of heat by a Fresnel's rhomb of rock-salt. These results have since been fully confirmed by the experiments of other observers.

EXAMPLES IN ACOUSTICS.

PERIOD, WAVE-LENGTH, AND VELOCITY.

1. If an undulation travels at the rate of 100 ft. per second, and the wave-length is 2 ft., find the period of vibration of a particle, and the number of vibrations which a particle makes per second.

2. It is observed that waves pass a given point once in every 5 seconds, and that the distance from crest to crest is 20 ft. Find the velocity of the waves in feet per second.

3. The lowest and highest notes of the normal human voice have about 80 and 800 vibrations respectively per second. Find their wave-lengths when the velocity of sound is 1100 ft. per second.

4. Find their wave-lengths in water in which the velocity of sound is 4900 feet per second.

5. Find the wave-length of a note of 500 vibrations per second in steel in which the velocity of propagation is 15,000 ft. per second.

PITCH AND MUSICAL INTERVALS.

6. Show that a "fifth" added to a "fourth" makes an octave.

7. Calling the successive notes of the gamut Do_1, Re_1, Mi_1, Fa_1, Sol_1, La_1, Si_1, Do_2, show that the interval from Sol_1 to Re_2 is a true "fifth."

8. Find the first 5 harmonics of Do_1.

9. A siren of 15 holes makes 2188 revolutions in a minute when in unison with a certain tuning-fork. Find the number of vibrations per second made by the fork.

10. A siren of 15 holes makes 440 revolutions in a quarter of a minute when in unison with a certain pipe. Find the note of the pipe (in vibrations per second).

REFLECTION OF SOUND, AND TONES OF PIPES.

11. Find the distance of an obstacle which sends back the echo of a sound to the source in $1\frac{1}{2}$ seconds, when the velocity of sound is 1100 ft. per second.

12. A well is 210 ft. deep to the surface of the water. What time will elapse between producing a sound at its mouth and hearing the echo?

13. What is the wave-length of the fundamental note of an open organ-pipe 16 ft. long; and what are the wave-lengths of its first two overtones? Find also their vibration-numbers per second.

14. What is the wave-length of the fundamental tone of a stopped organ pipe

5 ft. long; and what are the wave-lengths of its first two overtones? Find also their vibration-numbers per second.

15. What should be the length of a tube stopped at one end that it may resound to the note of a tuning-fork which makes 520 vibrations per second; and what should be the length of a tube open at both ends that it may resound to the same fork. [The tubes are supposed narrow, and the smallest length that will suffice is intended.]

16. Would tubes twice as long as those found in last question resound to the fork? Would tubes three times as long?

BEATS.

17. One fork makes 256 vibrations per second, and another makes 260. How many beats will they give in a second when sounding together?

18. Two sounds, each consisting of a fundamental tone with its first two harmonics, reach the ear together. One of the fundamental tones has 300 and the other 302 vibrations per second. How many beats per second are due to the fundamental tones, how many to the first harmonics, and how many to the second harmonics?

19. A note of 225 vibrations per second, and another of 336 vibrations per second, are sounded together. Each of the two notes contains the first two harmonics of the fundamental. Show that two of the harmonics will yield beats at the rate of 3 per second.

VELOCITY OF SOUND IN GASES.

20. If the velocity of sound in air at 0° C. is 33,240 cm. per second, find its velocity in air at 10° C., and in air at 100° C.

21. If the velocity of sound in air at 0° C. is 1090 ft. per second, what is the velocity in air at 10°?

22. Show that the difference of velocity for 1° of difference of temperature in the Fahrenheit scale is about 1 ft. per second.

23. If the wave-length of a certain note be 1 metre in air at 0°, what is it in air at 10°?

24. The density of hydrogen being ·06926 of that of air at the same pressure and temperature, find the velocity of sound in hydrogen at a temperature at which the velocity in air is 1100 ft. per second.

25. The quotient of pressure (in dynes per sq. cm.) by density (in gm. per cubic cm.) for nitrogen at 0° C. is 807 million. Compute (in cm. per second) the velocity of sound in nitrogen at this temperature.

26. If a pipe gives a note of 512 vibrations per second in air, what note will it give in hydrogen?

27. A pipe gives a note of 100 vibrations per second at the temperature 10° C. What must be the temperature of the air that the same pipe may yield a note higher by a major fifth?

VIBRATIONS OF STRINGS.

28. Find, in cm. per second, the velocity with which pulses travel along a string whose mass per cm. of length is ·005 gm., when stretched with a force of 7 million dynes.

29. If the length of the string in last question be 33 cm., find the number of vibrations that it makes per second when vibrating in its fundamental mode; also the numbers corresponding to its first two overtones.

30. The A string of a violin is 33 cm. long, has a mass of ·0065 gm. per cm., and makes 440 vibrations per second. Find the stretching force in dynes.

31. The E string of a violin is 33 cm. long, has a mass of ·004 gm. per cm., and makes 660 vibrations per second. Find the stretching force in dynes.

32. Two strings of the same length and section are formed of materials whose specific gravities are respectively d and d'. Each of these strings is stretched with a weight equal to 1000 times its own weight. What is the musical interval between the notes which they will yield?

33. The specific gravity of platinum being taken as 22, and that of iron as 7·8, what must be the ratio of the lengths of two wires, one of platinum and the other of iron, both of the same section, that they may vibrate in unison when stretched with equal forces?

Longitudinal Vibrations of Rods.

34. If sound travels along fir in the direction of the fibres at the rate of 15,000 ft. per second, what must be the length of a fir rod that, when vibrating longitudinally in its fundamental mode, it may emit a note of 750 vibrations per second?

35. A rod 8 ft. long, vibrating longitudinally in its fundamental mode, gives a note of 800 vibrations per second. Find the velocity with which pulses are propagated along it.

EXAMPLES IN OPTICS.

Photometry, Shadows, and Plane Mirrors.

36. A lamp and a taper are at a distance of 4·15 m. from each other; and it is known that their illuminating powers are as 6 to 1. At what distance from the lamp, in the straight line joining the flames, must a screen be placed that it may be equally illuminated by them both?

37. Two parallel plane mirrors face each other at a distance of 3 ft., and a small object is placed between them at a distance of 1 ft. from the first mirror, and therefore of 2 ft. from the second. Calculate the distances, from the first mirror, of the three nearest images which are seen in it; and make a similar calculation for the second mirror.

38. Show that a person standing upright in front of a vertical plane mirror will just be able to see his feet in it, if the top of the mirror is on a level with his eyes, and its height from top to bottom is half the height of his eyes above his feet.

39. A square plane mirror hangs exactly in the centre of one of the walls of a cubical room. What must be the size of the mirror that an observer with his

eye exactly in the centre of the room may just be able to see the whole of the opposite wall reflected in it except the part concealed by his body?

40. Two plane mirrors contain an angle of 160°, and form images of a small object between them. Show that if the object be within 20° of either mirror there will be three images; and that if it be more than 20° from both, there will be only two.

41. Show that when the sun is shining obliquely on a plane mirror, an object directly in front of the mirror may give two shadows, besides the direct shadow.

42. A person standing beside a river near a bridge observes that the inverted image of the concavity of the arch receives his shadow exactly as a real inverted arch would do if it were in the place where the image appears to be. Explain this.

43. If a globe be placed upon a table, show that the breadth of the elliptic shadow cast by a candle (considered as a luminous point) will be independent of the position of the globe.

44. What is the length of the cone of the *umbra* thrown by the earth? and what is the diameter of a cross section of it made at a distance equal to that of the moon?

The radius of the sun is 112 radii of the earth; the distance of the moon from the earth is 60 radii of the earth; and the distance of the sun from the earth is 24,000 radii of the earth. Atmospheric refraction is to be neglected.

45. The stem of a siren carries a plane mirror, thin, polished on both sides, and parallel to the axis of the stem. The siren gives a note of 345 vibrations per second. The revolving plate has 15 holes. A fixed source of light sends to the mirror a horizontal pencil of parallel rays. What space is traversed in a second by a point of the reflected pencil at a distance of 4 metres from the axis of the siren? This axis is supposed vertical.

Spherical Mirrors.

46. Find the focal length of a concave mirror whose radius of curvature is 2 ft., and find the position of the image (*a*) of a point 15 in. in front of the mirror; (*b*) of a point 10 ft. in front of the mirror; (*c*) of a point 9 in. in front of the mirror; (*d*) of a point 1 in. in front of the mirror.

47. Calling the diameter of the object unity, find the diameters of the image in the four preceding cases.

48. The flame of a candle is placed on the axis of a concave spherical mirror at the distance of 154 cm., and its image is formed at the distance of 45 cm. What is the radius of curvature of the mirror?

49. On the axis of a concave spherical mirror of 1 m. radius, an object 9 cm. high is placed at a distance of 2 m. Find the size and position of the image.

50. What is the size of the circular image of the sun which is formed at the principal focus of a mirror of 20 m. radius? The apparent diameter of the sun is 30′.

51. In front of a concave spherical mirror of 2 metres' radius is placed a concave luminous arrow, 1 decimetre long, perpendicular to the principal axis, and at the distance of 5 metres from the mirror. What are the position and size of the image? A small plane reflector is then placed at the principal focus of the spherical mirror, at an inclination of 45° to the principal axis, its polished side being next the mirror. What will be the new position of the image?

Refraction.

(The index of refraction of glass is to be taken as $\frac{3}{2}$, except where otherwise specified, and the index of refraction of water as $\frac{4}{3}$).

52. The sine of 45° is $\sqrt{\frac{1}{2}}$, or ·707 nearly. Hence, determine whether a ray incident in water at an angle of 45° with the surface will emerge or will be reflected; and determine the same question for a ray in glass.

53. If the index of refraction from air into crown-glass be $1\frac{1}{2}$, and from air into flint-glass $1\frac{2}{3}$, find the index of refraction from crown-glass into flint-glass.

54. The index of refraction from water into oil of turpentine is 1·11; find the index of refraction from air into oil of turpentine.

55. The index of refraction for a certain glass prism is 1·6, and the angle of the prism is 10°. Find approximately the deviation of a ray refracted through it nearly symmetrically.

56. A ray of light falls perpendicularly on the surface of an equilateral prism of glass with a refracting angle of 60°. What will be the deviation produced by the prism? Index of refraction of glass 1·5.

57. A speck in the interior of a piece of plate-glass appears to an observer looking normally into the glass to be 2 mm. from the near surface. What is its real distance?

58. The rays of a vertical sun are brought to a focus by a lens at a distance of 1 ft. from the lens. If the lens is held just above a smooth and deep pool of water, at what depth in the water will the rays come to a focus?

59. A mass of glass is bounded by a convex surface, and parallel rays incident nearly normally on this surface come to a focus in the interior of the glass at a distance a. Find the focal length of a plano-convex lens of the same convexity, supposing the rays to be incident on the convex side.

60. Show that the deviation of a ray going through an air-prism in water is towards the edge of the prism.

Lenses, &c.

61. Compare the focal lengths of two lenses of the same size and shape, one of glass and the other of diamond, their indices of refraction being respectively 1·6 and 2·6.

62. If the index of refraction of glass be $\frac{3}{2}$, show that the focal length of an equi-convex glass lens is the same as the radius of curvature of either face.

63. The focal length of a convex lens is 1 ft. Find the positions of the image of a small object when the distances of the object from the lens are respectively 20 ft., 2 ft., and $1\frac{1}{2}$ ft. Are the images real or virtual?

64. When the distances of the object from the lens in last question are respectively 11 in., 10 in., and 1 in., find the distances of the image. Are the images real or virtual?

65. Calling the diameter of the object unity, find the diameter of the image in the six cases of questions 63, 64, taken in order.

66. Show that, when the distance of an object from a convex lens is double the focal length, the image is at the same distance on the other side.

67. The object is 6 ft. on one side of a lens, and the image is 1 ft. on the other side. What is the focal length of the lens?

68. The object is 3 in. from a lens, and its image is 18 in. from the lens on the same side. Is the lens convex or concave, and what is its focal length?

69. The object is 12 ft. from a lens, and the image 1 ft. from the lens on the same side. Find the focal length, and determine whether the lens is convex or concave.

70. A person who sees best at the distance of 3 ft., employs convex spectacles with a focal length of 1 ft. At what distance should he hold a book, to read it with the aid of these spectacles?

71. A person reads a book at the distance of 1 ft. with the aid of concave spectacles of 6 in. focal length. At what distance is the image which he sees?

72. A pencil of parallel rays fall upon a sphere of glass of 1 inch radius. Find the principal focus of rays near the axis, the index of refraction of glass being 1·5.

73. What is the focal length of a double-convex lens of diamond, the radius of curvature of each of its faces being 4 millimetres? Index of refraction 2·5.

74. An object 8 centimetres high is placed at 1 metre distance on the axis of an equi-convex lens of crown-glass of index 1·5, the radius of curvature of its faces being 0·4 m. Find the size and position of the image.

75. Two converging lenses, with a common focal length of 0·05 m., are at a distance of 0·05 m. apart, and their axes coincide. What image will this system give of a circle 0·01 m. in diameter, placed at a distance of ·1 m. on the prolongation of the common axis?

76. Show that if F denote the focal length of a combination of two lenses in contact, their thicknesses being neglected, we have

$$\frac{1}{F} = \frac{1}{f_1} + \frac{1}{f_2},$$

f_1 and f_2 denoting the focal lengths of the two lenses.

77. What is the focal length of a lens composed of a convex lens of 2 in. focal length, cemented to a concave lens of 9 in. focal length?

78. Apply the formulæ of § 1015 to find the focal length of a lens, the thickness being neglected.

79. The objective of a telescope has a focal length of 20 ft. What will be the magnifying power when an eye-piece of half-inch focus is used?

80. A sphere of glass of index 1·5 lying upon a horizontal plane receives the sun's rays. What must be the height of the sun above the horizon that the principal focus of the sphere may be in this horizontal plane?

81. A small plane mirror is placed exactly at the principal focus of a telescope, nearly perpendicular to its axis, and the telescope is directed approximately to a distant luminous object. Show that the rays reflected at the mirror will, after repassing the object glass, return in the exact direction from which they came, in spite of the small errors of adjustment of the mirror and telescope.

82. An eye is placed close to the surface of a large sphere of glass ($\mu=\frac{3}{2}$) which is silvered at the back. Show that the image which the eye sees of itself is three-fifths of the natural size.

83. The refractive indices for the rays D and F for two specimens of glass are

Crown-glass	......	1·5279	......	1·5344
Flint-glass	......	1·6351	......	1·6481

and an achromatic lens of 20 in. focal length is to be formed by their combination. Show that if the rays D and F are brought to the same focus, the focal lengths of the two lenses which are combined will be about 7·9 in. for the crown and 13·1 in. for the flint.

ANSWERS TO EXAMPLES IN ACOUSTICS.

Ex. 1. $\frac{1}{50}$ sec. 50. Ex. 2. 4. Ex. 3. $13\frac{3}{4}$ ft. $1\frac{3}{8}$ ft. Ex. 4. $61\frac{1}{4}$ ft., $6\frac{1}{3}$ ft. Ex. 5. 30 ft.

Ex. 6. $\frac{3}{2} \times \frac{4}{3} = 2$. Ex. 7. $\frac{36}{54} = \frac{2}{3}$. Ex. 8. Do_2, Sol_2, Do_3, Mi_3, Sol_3.

Ex. 9. 547. Ex. 10. 440.

Ex. 11. 825. Ex. 12. $\frac{21}{55} = {\cdot}382$ sec. Ex. 13. 32 ft., 16 ft., $10\frac{2}{3}$ ft.; $34\frac{3}{8}$, $68\frac{3}{4}$, $103\frac{1}{8}$. Ex. 14. 20 ft., $6\frac{2}{3}$ ft., 4 ft.; 55, 165, 275. Ex. 15. $\frac{55}{104}$ ft., $\frac{55}{52}$ ft. Ex. 16. An open tube twice or three times as long will resound, because one of its overtones will coincide with the note of the fork. A stopped tube three times as long will resound, but a stopped tube twice as long will not.

Ex. 17. 4. Ex. 18. 2, 4, 6. Ex. 19. $675 - 672 = 3$.

Ex. 20. 33843, 38850. Ex. 21. 1110. Ex. 22. The velocity is 1090 at 32° and 1110 at 50°. Ex. 23. 1·018 metre. Ex. 24. 4180 ft. per second. Ex. 25. 33732. Ex. 26. 1945 vibrations per second. Ex. 27. 364° C.

Ex. 28. 37417. Ex. 29. 567, 1134, 1701. Ex. 30. $v = 29040$, $t = v^2 m = 5481600$. Ex. 31. $v = 43560$, $t = 7589900$. Ex. 32. Unison. Ex. 33. Length of iron $= 1{\cdot}68$ times length of platinum.

Ex. 34. 10 ft. Ex. 35. 12800 ft. per second.

ANSWERS TO EXAMPLES IN OPTICS.

Ex. 36. 2·95 m. Ex. 37. 1, 5, and 7 ft. behind first mirror; 2, 4, and 8 ft. behind second. Ex. 39. Side of mirror must be $\frac{1}{3}$ of edge of cube.

Ex. 41. They are the shadows of the object and of its image, cast by the sun's image. The former is due to the intercepting of light after reflection; the latter to the intercepting of light before reflection. Ex. 42. The sun's image throws a shadow of the man's image on the real arch, owing to his intercepting rays on their way to the water. Ex. 43. First let the globe be vertically under the flame, and draw through the flame two equally inclined planes, touching the globe. Their intersections with the table will be parallel lines which will be tangents to the shadow, and will still remain tangents to it as the globe is rolled between the planes to any distance. Ex. 44. 216 radii of earth; $1\frac{4}{9}$ radii. Ex. 45. $368\pi = 1156$ metres.

Ex. 46. Focal length 1 ft.; (*a*) 5 ft. in front of mirror; (*b*) $1\frac{1}{5}$ ft. in front; (*c*) 3 ft. behind mirror; (*d*) $1\frac{1}{11}$ in. behind. Ex. 47. 4, $\frac{1}{5}$, 4, $1\frac{1}{11}$.

Ex. 48. 69·6 cm. Ex. 49. Distance $\frac{2}{3}$ m., height 3 cm. Ex. 50. 8·73 cm. Ex. 51. Distance $1\frac{1}{4}$ m., length $\frac{1}{4}$ dec., new position $\frac{1}{4}$ m. laterally from focus.

Ex. 52. The ray in water will emerge, because $\frac{3}{4}$ is greater than ·707; the ray

in glass will be totally reflected, because $\frac{2}{3}$ is less than ·707. Ex. 53. $\frac{16}{15}$.
Ex. 54. 1·48. Ex. 55. 6°. Ex. 56. 60° (by total reflection).

Ex. 57. 3 mm. Ex. 58. 1 ft. 4 in. Ex. 59. $\frac{2}{3}a$.

Ex. 61. Focal length of diamond lens is $\frac{3}{8}$ of focal length of glass lens.
Ex. 63. $1\frac{1}{19}$ ft., 2 ft., 3 ft. on other side of lens. All real. Ex. 64. 11 ft., 5 ft.,
$\frac{1}{11}$ ft. on same side of lens. All virtual. Ex. 65. $\frac{1}{19}$, 1, 2, 12, 6, $1\frac{1}{11}$. Ex. 66. $1\frac{1}{11}$.
Ex. 67. $\frac{4}{7}$ ft. Ex. 68. $3\frac{3}{5}$ in., convex. Ex. 69. $1\frac{1}{11}$ ft., concave. Ex. 70. 9 in.
Ex. 71. 4 in. Ex. 72. 1·5 in. from centre, or ·5 in. from sphere.

Ex. 73. $1\frac{1}{3}$ mm. Ex. 74. Distance $\frac{2}{3}$ m. on other side, height $5\frac{1}{3}$ cm. Ex. 75. A
real image ·025 m. beyond second lens; diameter of image ·005 m. Ex. 77. $2\frac{4}{7}$ in.
Ex. 79. 480. Ex. 80. Sine of altitude $=\frac{2}{3}$, altitude $=41° \, 49'$. Ex. 81. Rays from
one point of object converge to one point on mirror, and are reflected from this
point as a new source. Hence by the principle of conjugate foci they will return
to the point whence they came. Ex. 82. The first and second images are at distances of $\frac{1}{3}$ and $\frac{2}{3}$ of radius from centre.

Ex. 83. The dispersive powers are as 32 : 53. The focal lengths are to be
directly as these numbers, and the difference of their reciprocals must be $\frac{1}{20}$.

INDEX TO PART IV.

INDEX TO PART IV.

THE END.

$a'OP + a'Oa = L$

$a'OP + POP' = L$

$\therefore a'Oa = POP'$

$POI = POR$

[illegible]

[illegible] $= 2O$

$IOP' = POP' + IOP = I'OP'$

[illegible] $+ (IOP + ROP'$

$I'O$ [illegible]

[illegible] $2a'Oa$

$IOR + ROP' = IOP'$

$I'OP' + ROP' = I'OR$

INTERNATIONAL SERIES IN PURE AND APPLIED PHYSICS

LEONARD I. SCHIFF, *Consulting Editor*

Allis and Herlin Thermodynamics and Statistical Mechanics
Becker Introduction to Theoretical Mechanics
Clark Applied X-rays
Edwards Analytic and Vector Mechanics
Evans The Atomic Nucleus
Finkelnburg Atomic Physics
Green Nuclear Physics
Gurney Introduction to Statistical Mechanics
Hall Introduction to Electron Microscopy
Hardy and Perrin The Principles of Optics
Harnwell Electricity and Electromagnetism
Harnwell and Livingood Experimental Atomic Physics
Harnwell and Stephens Atomic Physics
Houston Principles of Mathematical Physics
Houston Principles of Quantum Mechanics
Hughes and DuBridge Photoelectric Phenomena
Hund High-frequency Measurements
Kemble The Fundamental Principles of Quantum Mechanics
Kennard Kinetic Theory of Gases
Marshak Meson Physics
Morse Vibration and Sound
Morse and Feshbach Methods of Theoretical Physics
Muskat Physical Principles of Oil Production
Read Dislocations in Crystals
Richtmyer, Kennard, and Lauritsen Introduction to Modern Physics
Schiff Quantum Mechanics
Seitz The Modern Theory of Solids
Slater Introduction to Chemical Physics
Slater Microwave Transmission
Slater Quantum Theory of Matter
Slater and Frank Electromagnetism
Slater and Frank Introduction to Theoretical Physics
Slater and Frank Mechanics
Smythe Static and Dynamic Electricity
Squire Low Temperature Physics
Stratton Electromagnetic Theory
Thorndike Mesons: A Summary of Experimental Facts
Townes and Schawlow Microwave Spectroscopy
White Introduction to Atomic Spectra

The late F. K. Richtmyer was Consulting Editor of the series from its inception in 1929 to his death in 1939. Lee A. DuBridge was Consulting Editor from 1939 to 1946; and G. P. Harnwell from 1947 to 1954.

Electromagnetism

by JOHN C. SLATER, PH.D.
Professor of Physics
Massachussetts Institute of Technology

and NATHANIEL H. FRANK, SC.D.
Professor of Physics
Massachusetts Institute of Technology

ASIAN STUDENTS' EDITION

NEW YORK AND LONDON
McGRAW-HILL BOOK COMPANY, INC.

TOKYO
KŌGAKUSHA COMPANY, LTD.

ELECTROMAGNETISM

ASIAN STUDENTS' EDITION

TOSHO INSATSU
PRINTING CO., LTD.
TOKYO. JAPAN

PREFACE

The present book is the second of several volumes which are intended to replace the *Introduction to Theoretical Physics* written by the same authors in 1933. By separating the material on mechanics, on electromagnetism, and on the quantum theory, we believe that it is possible to give a somewhat better rounded treatment of each of these fields, which will be more useful to the teacher and the student. We have taken advantage of the opportunity to give a considerably more complete treatment of the foundations of electrostatics and magnetostatics, and to introduce some of the new developments in electromagnetic theory since 1933. At the same time, we have tried to preserve, and even to extend, the general unity of treatment which we believed so important when the earlier book was published. We have a conviction that the teaching of theoretical physics in a number of separate courses, as in mechanics, electromagnetic theory, potential theory, thermodynamics, and modern physics, tends to keep a student from seeing the unity of physics, and from appreciating the importance of applying principles developed for one branch of science to the problems of another.

We have developed electromagnetism from first principles, and have included in the appendixes enough of the necessary mathematical background so that the student familiar with the calculus and differential equations can follow the work, without further training in mathematics or in mathematical physics. Nevertheless, electromagnetic theory is a later historical development than mechanics, and it makes use of many mathematical methods, as in vector and tensor analysis, potential theory, and partial differential equations, which were first developed for mechanics, and find their most straightforward applications in mechanical problems. Although we feel that the present volume is designed so that it can be used by itself, a thorough grounding in mechanics is a very desirable prerequisite. The student who first familiarizes himself with the companion volume on mechanics will find many similarities in treatment of the two fields, which will enhance his understanding of both.

Electromagnetic theory has developed in two principal directions since its original formulation in the middle of the last century: toward

the electromagnetic theory of light, treating very short wave lengths, and providing the mathematical foundation for the theory of optics; and toward the longer wave lengths encountered in electrical engineering and radio. In the last few years, the practical applications have come at constantly shorter wave lengths, until with the recent development of microwaves the gap between the two branches of the subject has been practically closed. We handle problems of both types, making no distinction between the methods used for them.

In 1933, it was almost universal practice to use the Gaussian system of units for handling electromagnetic problems, and *Introduction to Theoretical Physics* is written in terms of those units. Since then, the mks, or meter-kilogram-second, system has come into common use, and we have made use of that system in the present volume, believing that its advantages over the Gaussian system are considerable. We give in an appendix, however, a discussion of the various systems of units, including a formulation of all the important equations in terms of the Gaussian system, so that the reader will not be at a disadvantage in consulting other books using the Gaussian units, and so that he will understand properly the relations among the various systems.

We quote from the preface of *Introduction to Theoretical Physics:*

"In a book of such wide scope, it is inevitable that many important subjects are treated in a cursory manner. An effort has been made to present enough of the groundwork of each subject so that not only is further work facilitated, but also the position of these subjects in a more general scheme of physical thought is clearly shown. In spite of this, however, the student will of course make much use of other references, and we give a list of references, by no means exhaustive, but suggesting a few titles in each field which a student who has mastered the material of this book should be able to appreciate.

"At the end of each chapter is a set of problems. The ability to work problems, in our opinion, is essential to a proper understanding of physics, and it is hoped that these problems will provide useful practice. At the same time, in many cases, the problems have been used to extend and amplify the discussion of the subject matter, where limitations of space made such discussion impossible in the text. The attempt has been made, though we are conscious of having fallen far short of succeeding in it, to carry each branch of the subject far enough so that definite calculations can be made with it. Thus a far surer mastery is attained than in a merely descriptive discussion.

"Finally, we wish to remind the reader that the book is very defi-

nitely one on theoretical physics. Though at times descriptive material, and descriptions of experimental results, are included, it is in general assumed that the reader has a fair knowledge of experimental physics, of the grade generally covered in intermediate college courses. No doubt it is unfortunate, in view of the unity which we have stressed, to separate the theoretical side of the subject from the experimental in this way. This is particularly true when one remembers that the greatest difficulty which the student has in mastering theoretical physics comes in learning how to apply mathematics to a physical situation, how to formulate a problem mathematically, rather than in solving the problem when it is once formulated. We have tried wherever possible, in problems and text, to bridge the gap between pure mathematics and experimental physics. But the only satisfactory answer to this difficulty is a broad training in which theoretical physics goes side by side with experimental physics and practical laboratory work. The same ability to overcome obstacles, the same ingenuity in devising one method of procedure when another fails, the same physical intuition leading one to perceive the answer to a problem through a mass of intervening detail, the same critical judgment leading one to distinguish right from wrong procedures, and to appraise results carefully on the ground of physical plausibility, are required in theoretical and in experimental physics. Leaks in vacuum systems or in electric circuits have their counterparts in the many disastrous things that can happen to equations. And it is often as hard to devise a mathematical system to deal with a difficult problem, without unjustifiable approximations and impossible complications, as it is to design apparatus for measuring a difficult quantity or detecting a new effect. These things cannot be taught. They come only from that combination of inherent insight and faithful practice which is necessary to the successful physicist. But half the battle is over if the student approaches theoretical physics, not as a set of mysterious formulas, or as a dull routine to be learned, but as a collection of methods, of tools, of apparatus, subject to the same sort of rules as other physical apparatus, and yielding physical results of great importance. . . . The aim has constantly been, not to teach a great collection of facts, but to teach mastery of the tools by which the facts have been discovered and by which future discoveries will be made."

John C. Slater
Nathaniel H. Frank

Cambridge, Mass.,
July, 1947.

CONTENTS

CHAPTER I

THE FIELD THEORY OF ELECTROMAGNETISM

A dynamical problem has two aspects: mechanics, the determination of the accelerations and hence of the motions, once the forces are given; and the study of the forces acting under the existing circumstances. The basic principles of mechanics are simple. In its classical form, mechanics is based on Newton's laws of motion, laws discovered and formulated nearly three hundred years ago. The developments since then have been technical, mathematical improvements in the way of formulating the laws and solving the resulting mathematical problems, rather than additions to our fundamental knowledge of mechanics. Only in the present century, with wave mechanics, has there been a change in the underlying structure of the subject.

The study of forces, on the other hand, is difficult and complex. The first forces brought into mathematical formulation were gravitational forces, as seen in planetary motion. Next were elastic forces. Then followed electric and magnetic forces, which are the subject of this volume. Their study was mostly a product of the nineteenth century. During the present century, it has become clear that electromagnetic forces are of far wider application than was first supposed. It has become evident that, instead of being active only in electrostatic and magnetostatic experiments, and in electromagnetic applications such as the telegraph, dynamo, and radio, the forces between the nuclei and electrons of single atoms, the chemical forces between atoms and molecules, the forces of cohesion and elasticity holding solids together, are all of an electric nature. We might be tempted to generalize and suppose that all forces are electromagnetic, but this appears to be carrying things too far. The prevailing evidence at present indicates that the intranuclear forces, holding together the various fundamental particles of which the nucleus is composed, are not of electromagnetic origin. These forces, of enormous magnitude, and appearing in the phenomena of radioactivity and of nuclear fission, appear subject to laws somewhat analogous to the electromagnetic laws, but fundamentally different. In spite of this, the range of phenomena governed by electromagnetic theory is very wide, and it carries us rather far into the structure of matter, of electrons and

nuclei and atoms and molecules, if we wish to understand it completely. The equations underlying the theory, Maxwell's equations, are relatively simple, but not nearly so simple as Newton's laws of motion. Instead of stating the whole fundamental formulation of the subject in the first chapter, as one can when dealing with mechanics, about half of the present book is taken up with a complete formulation of Maxwell's equations. We start with simple types of force, electrostatic and magnetostatic, and gradually work up to problems of electromagnetic induction and related subjects, all of which are formulated in Maxwell's equations.

In the development of electromagnetic theory, there has been a continual and significant trend, which in a way has set the pattern for the development of all of theoretical physics. This has been the trend away from the concept known as "action at a distance" toward the concept of field theory. The classical example of action at a distance is gravitation, in which simple nonrelativistic theory states that any two particles in the universe exert a gravitational force on each other, acting along the line joining them, proportional to the product of their masses, and inversely proportional to the square of the distance between them. Such a force, depending only on the positions of the particles, quite independent of any intervening objects, is simple to think about, and formed the basis of most of physical thought from the time of Newton, in the latter half of the seventeenth century, on well into the nineteenth century. The first electric and magnetic laws to be discovered fitted in well with the pattern. First among these was Coulomb's law. Coulomb investigated the forces between electrically charged objects, and found that the force between two such objects was in the line joining them, proportional to the product of their charges (which could be defined by an experiment based on this observation), and inversely proportional to the square of the distance between them, in striking analogy to the law of gravitation. Magnets similarly fell in with the scheme. A theory of the forces between permanent magnets can be built up by considering that they contain magnetic north and south poles, and that the force between two poles is proportional to the product of the pole strengths, and inversely proportional to the square of the distance. It is true that single poles do not seem to exist in nature, but an ordinary magnet can be considered as made up of equal north and south poles in juxtaposition, a combination known as a "dipole."

Coulomb's studies were carried on in the latter half of the eighteenth century. Early in the nineteenth came the discovery of the

magnetic effects of continuous currents. First was Oersted's observation that electric current flowing in a loop of wire exerted magnetic forces on permanent magnets, just as if the loop itself were a magnet. Then came Ampère to formulate these observations mathematically, showing that the magnetic force resulting from a circuit can be broken up into contributions from infinitesimal lengths of wire in the circuit, and that each of these forces falls off as the inverse square of the distance, a law often known by the names of Biot and Savart. These laws of action at a distance suggested that electromagnetism would develop along the lines suggested by gravitational theory.

Michael Faraday, in the first half of the nineteenth century, was the first who really turned the electromagnetic theory into the lines of field theory. If a piece of insulator, or dielectric, is put between two charged objects, the force between the objects is diminished. Faraday was not content to regard this as merely a shielding effect, or a change in the force constant. He directed attention to the dielectric, and concluded that it became polarized, acquired charges which themselves contributed to the force on other charges. To describe these effects, he introduced the idea of lines of force, lines pointing in the direction of the force that would be exerted on a charge located at an arbitrary point of space. He gave a physical meaning to the number of lines per unit area, setting this quantity proportional to the magnitude of the force. He thought of the lines of force in a very concrete way, as if there were a tension exerted along them, and a pressure at right angles to them, and showed that such a stress system would account for the forces actually exerted on charges. Faraday's fundamental idea, in other words, was that things of the greatest importance were going on in the apparently empty space between charged bodies, and that electromagnetism could be described by giving the laws of the phenomena in this space, which he called the "field." His discovery of electromagnetic induction, in which electromotive force is induced in a circuit by the time rate of change of magnetic flux through the circuit, added certainty to his concepts, by pointing out the importance of the magnetic field and its flux.

Faraday was not a mathematician, and his concepts of the field did not immediately appeal to the mathematicians, who were still thinking in terms of inverse-square laws. His contemporary Gauss furnished the first mathematical formulation of field theory. Gauss considered lines of force, their flux out of a region, and proved his famous theorem, relating this flux to the total charge within the region. It remained for Maxwell, however, some thirty years after

Faraday's first discoveries, to find the real mathematical formulation of them. Maxwell accepted wholeheartedly the idea that the electric and magnetic fields were the fundamental entities, and considered the partial differential equations governing those fields. He had a background of experience to work on. In addition to the work of Gauss, there was the formulation of gravitational theory in terms of the gravitational field and potential, which had been worked out at the end of the eighteenth century by Laplace, Poisson, and others. At the time, that formulation seemed more a mathematical device than anything else, but in the hands of Maxwell it furnished an ideal mathematical framework for Faraday's ideas. The electromagnetic field is much more complicated than the gravitational, however, and Maxwell had to go far beyond Laplace, Poisson, and Gauss, introducing among other things the concept of displacement current, which proved to be necessary to reach a mathematically consistent theory. Maxwell's equations have stood the test of time since then, and still furnish the correct formulation of classical electromagnetic theory; it is only the quantum theory which has brought about a fundamental revision of our ideas, during the last few years.

As soon as Maxwell formulated his equations, he was able to draw from them a mathematical result predicting a new phenomenon, which would hardly be suspected from the laws of Coulomb and of Faraday which were his starting points. He was able to show that an electromagnetic disturbance originated by one charged body would not be immediately observed by another, but that instead it would travel out as a wave, with a speed that could be predicted from electrical and magnetic measurements. Furthermore, the velocity so predicted proved to agree, within the small experimental error, with the speed of light. Thus at one blow he accomplished two results of the greatest importance in the history of physics. First, he gave a convincing proof of the superiority of a field theory to action at a distance; secondly, he tied together two great branches of physics, electromagnetism and optics.

To see why action at a distance can hardly explain the propagation of electromagnetic waves, consider as simple a thing as a radio broadcast. In the transmitting antenna, certain charges oscillate back and forth, depending on the signal being transmitted. According to the field theory, these charges produce an electromagnetic wave, which travels out with the speed of light. The wave reaches a receiving antenna an appreciable time later, and sets the charges in that antenna into oscillation, with results that can be detected in the

receiver. The forces on the charges in the receiving antenna are not determined at all by the instantaneous positions or velocities of the charges in the transmitting antenna, but by the values that they had at an earlier time. Any reaction back on the transmitter will be delayed by the time taken by the disturbance to reach the receiver, and then to return to the transmitter again, as in an echo. The forces on a particle, in other words, do not depend on the positions of other charges, but on what they did at past times. It is almost impossible to formulate this in terms of action at a distance, but easy to formulate if we regard the electromagnetic field as a real entity, taking energy from the transmitter, and carrying it with a finite velocity to the receiver.

To appreciate the relations between electromagnetism and optics, which Maxwell demonstrated, we have to go back somewhat further with the development of optics. At the time of Newton and Huygens, there were two opposed theories of light, Newton holding a corpuscular theory, in which the light was a stream of infinitesimal particles, being bent as they passed from one medium to another on account of a surface force resulting from different potential energies in the various media; whereas Huygens believed that light was a form of wave motion, and was able to explain reflection and refraction on the basis of the propagation of spherical wavelets, traveling with different velocities in different media. Newton's principal objection to the wave theory was his feeling that it did not explain the way in which obstacles cast sharp shadows. He was thinking by analogy with sound, which was known to be a wave motion, in which sound bends around obstacles. The thing he did not realize was that the wave lengths of light are so small, and that that entirely changes the behavior of shadow formation. It is curious that he did not think of this, for he was familiar with the phenomena of interference and diffraction; he made a theory of them which postulated a periodic disturbance along the direction of wave propagation, which he described as alternate fits of easy reflection and of easy transmission, and by measurement of interference patterns he determined the wave length of this periodic disturbance, in good agreement with modern measurements of the wave length of light. His combination of corpuscles with a periodic disturbance, in fact, showed extraordinary similarity to the present picture resulting from the quantum theory, in which we picture particles, or photons, traveling in accordance with a guiding wave field.

The explanation of interference and diffraction from the wave

theory, together with the proper treatment of the casting of shadows, did not actually come until the first years of the nineteenth century, when Young and Fresnel made their discoveries in that field. Fresnel not only explained these phenomena, but also formulated the laws of reflection and refraction, giving laws, which have proved to be correct, for the fraction of the incident light reflected and refracted, as well as for the direction of the reflected and refracted beams. Those laws also explained the phenomenon of double refraction, by which certain crystals such as Iceland spar transmit light in two different rays, the ordinary and the extraordinary rays, traveling at different angles and speeds, and not satisfying the ordinary laws of refraction as found in an isotropic medium. The two rays show properties described as polarization, which proved to result from the fact that light is a transverse, not a longitudinal, vibration, so that two directions of vibration are possible, both at right angles to the direction of propagation. The whole theory of these vibrations shows a close analogy to the transmission of transverse elastic vibrations in an elastic solid, with the one exception that there is no indication of an accompanying longitudinal vibration, such as there would be with an elastic solid, and such as would constitute the only mode of vibration for a fluid.

The physicists of the nineteenth century were much devoted to mechanical models. If light acted like the transverse vibration of an elastic solid, they tried to visualize it as a real solid, and gave it a name, the "ether." It was hard to understand its properties. In the first place, as we have just mentioned, a real elastic solid would transmit longitudinal as well as transverse waves, and no good way of modifying the theory was found that would eliminate the longitudinal waves. In the second place, the solid would have to fill all space, and it was obviously very hard to see how, with a very rigid solid filling space, it was possible for ordinary bodies to move around freely. Much thought was devoted to these questions. Even after Maxwell had shown that light was an electromagnetic disturbance, not a vibration of a solid, there was still much speculation about the nature of the ether. It is really only within the present century that physicists have realized that that speculation is essentially meaningless, that the electric and magnetic fields are the fundamental entities concerned with optics as well as with electromagnetic forces, and that we do not have to endow these fields with mechanical properties foreign to their real nature.

It was this background of an elastic-solid theory of light which

Maxwell encountered when he formulated his electromagnetic theory. That theory at once removed all the difficulties of the previous theories. It yielded only the transverse vibrations, having no solutions of the fundamental equations corresponding to longitudinal waves. Fresnel's equations for refraction and reflection, and his explanations of interference and diffraction, though proposed for an elastic-solid theory, proved to be equally valid in the framework of electromagnetic theory. And the explanation of the casting of shadows, resulting from the small wave length shown to exist by experiments on interference and diffraction, properly answered Newton's objection to the wave theory. Taken together with the remarkable success of the theory in predicting the velocity of light from purely electrical measurements, all doubt about the electromagnetic nature of light almost immediately disappeared, and optics is now treated as a branch of electromagnetism, as we shall treat it in this volume. The argument was placed beyond question a few years later, when Hertz demonstrated the existence of electromagnetic waves of wave lengths of a few centimeters. This was soon followed by the use of much longer electromagnetic waves for radio communication, leading back in the last few years to the use of microwaves, of a few centimeters in length, forming one of the most perfect examples of the application of Maxwell's equations.

The developments we have described comprise most of the classical part of electromagnetic theory, the part that is taken up in the present volume. Two more recent advances have concerned themselves with the quantum theory. First, and more revolutionary, has been the discovery of the photon, and of the fact that light has an aspect that must be explained in a corpuscular way, as well as a wave aspect. This was at first a mathematical deduction by Planck, from the theory of black-body radiation, at the beginning of the present century, and it was followed shortly by Einstein's deduction of the law of photoelectric emission, from Planck's quantum hypothesis. Light of rather short wave length, falling on a suitable metal surface, ejects electrons; these electrons are observed to have an energy depending only on the frequency of the light, being in fact proportional to the frequency, and being independent of the intensity of the light, which regulates merely the number of photoelectrons, not their energy. This suggested strongly that the energy of the light was carried by certain corpuscles, or photons, shown by Einstein to carry a quite finite energy hf, where h is Planck's constant, f the frequency, and that a photon absorbed by an electron of the metal transferred its energy

to the electron, so that it was ejected as a photoelectron. Such a hypothesis was in direct conflict with the wave theory, which as we shall see predicts a continuous distribution of energy, and for a number of years this conflict was regarded as a great stumbling block in the development of the quantum theory.

The difficulty still exists, in its way, but it is gradually being worked out in the framework of developing quantum electrodynamics. It is becoming recognized that the corpuscular and the wave point of view simultaneously have their truth, and that the wave is the correct description of the average behavior of the photons, but does not predict the behavior of an individual photon, which in fact seems not to be predictable by any precise theory. The photons move so that on the average they deliver the same energy to any illuminated area that would be predicted by the wave theory, but the energy is actually delivered by the photons, in discrete amounts. At comparatively long wave lengths, or low frequencies, the energy of the photons is so small that a great many are delivered by a source of ordinary energy, and their discreteness is not of importance. At very high frequencies, however, the finite size of the photons is of striking significance, as can be seen in experiments with counters and cloud chambers, in which evidence of a single photon (or of a single electron or ion) can be directly observed.

The less striking, but probably equally important, development of electromagnetic theory in the last few years has been its place in the quantum theory of atomic and molecular structure. Bohr, in 1913, was able to explain the structure of a hydrogen atom, as being made up of a proton and an electron, acting on each other by ordinary electrostatic forces, but governed by quantum mechanics. Developing this theory, Schrödinger in 1926 proposed his wave mechanics, a form of quantum mechanics suggested by the duality between corpuscles and waves in the theory of light. It had been suggested by de Broglie that there was a wave phenomenon associated with the motion of electrons, as there was with the motion of photons, and Schrödinger formulated this wave in terms of a wave equation, somewhat similar to the wave equations of optics. By solving this equation, we find descriptions of great accuracy of the structure of atoms, molecules, and solids and matter of all types. The interesting point is that the forces that enter into Schrödinger's theory are essentially electromagnetic forces as described by classical electromagnetic theory; it is in the mechanics that this theory differs from purely classical theory. Thus we understand the very broad way in which

electromagnetic forces underlie a great segment of chemistry, and of the theory of the solid, liquid, and gaseous states.

The applications of Schrödinger's theory to the structure of matter deal almost entirely with particles traveling slowly compared with the velocity of light. The electromagnetic theory that has to be used is then essentially electrostatics, making no use of the finite velocity of propagation of light. When we come to apply the theory to particles traveling with velocities comparable with that of light, we meet essential difficulties, which have not yet been fully solved. We must expect a combination of the theory of radiation, and of the electromagnetic forces concerned in an atom or molecule, which will be consistent and complete, and this does not yet seem to exist. Of course, particles traveling with velocities near that of light must satisfy the relativistic mechanics. Einstein, in the early years of the century, devised his theory of relativity, to explain the dynamics of rapidly moving particles, and the way in which the velocity of light appeared the same in any frame of reference. The modifications that proved to be necessary were only in the dynamics, not in the electromagnetic part of the theory; the forces as predicted by classical electromagnetic theory are relativistically correct, as had been shown earlier by Lorentz. But, in setting up the quantum electrodynamics of rapidly moving particles, it is relativistic rather than Newtonian mechanics which must be used as a start, and it is this combination which is not yet without its difficulties.

Probably part of the trouble with these theories is the fact that new phenomena are appearing as we go to very rapid particles of very high energy, which complicate the whole theory. These are the phenomena of the forces between the particles, neutrons and protons, in the nucleus. There is good evidence that these forces are not electromagnetic, but of another variety, produced not by the electromagnetic field, but by a meson field. This field has been described mathematically by analogy with electromagnetic theory. It similarly has a wave aspect, but also a corpuscular aspect, the corpuscles in this case being mesons, rather than photons. It is thus becoming likely that there exist in nature a number of different levels of forces and particles: the electromagnetic field and the photons, the electrons and protons and neutrons and their associated fields of the de Broglie or Schrödinger type, and the mesons and their fields. And the guiding pattern for the development of all these theories is at present the electromagnetic theory in its classical form. It remains to be seen what final synthesis of these various theories can eventually be made.

We are now ready to proceed with our study of classical electromagnetic theory, which alone we shall take up in this volume. We start in the historical order, treating electrostatics, first from the basis of Coulomb's law, then from the standpoint of field theory, showing as Faraday did how much that helps us in the study of problems involving dielectrics and conductors. Then we take up in a similar way the magnetic field, passing on through the study of electromagnetic induction to Maxwell's equations, and to their application to the electromagnetic theory of light and of other electromagnetic radiation fields.

1. The Force on a Charge.—Electromagnetic theory deals with the forces acting on charges and currents. We find that an electric charge at a given point of space is acted on by two types of force: an electric force, independent of its velocity, and a magnetic force, proportional to its velocity (that is, to the current carried by the charge), and at right angles to its velocity. We find that different charges at the same point of space are acted on by different amounts of force, and we arbitrarily define the strength of the charge as being proportional to the magnitude of force acting on it in a given field. We shall see later how to define the unit of charge, the coulomb. We can define two vectors at every point of space, $\mathbf{E}$ the electric intensity, $\mathbf{B}$ the magnetic induction, such that the force $\mathbf{F}$ on a charge of q coulombs moving with velocity $\mathbf{v}$ is given by the vector equation

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{1.1}$$

Here $\mathbf{v} \times \mathbf{B}$ is the vector product of $\mathbf{v}$ and $\mathbf{B}$, a vector at right angles to both, whose magnitude is the product of the magnitudes of $\mathbf{v}$ and $\mathbf{B}$ times the sine of the angle between. The reader unfamiliar with this and other aspects of vector notation will find vector methods discussed in Appendix I. To measure $\mathbf{E}$ and $\mathbf{B}$ we need only measure the force on a moving charge at the point in question. The second term is the ordinary motor rule, that the force on a current element is proportional to the current, and the component of magnetic field at right angles to it, and is at right angles to each. In the mks (meter-kilogram-second) system of units, which we shall use, the force will be given by (1.1) in newtons (1 newton = 10^5 dynes), if the charge q is in coulombs, if $\mathbf{E}$ is in volts per meter, $\mathbf{v}$ in meters per second, and $\mathbf{B}$ in webers per square meter (1 weber/sq m = 10^4 gausses). The mks system of units is discussed in Appendix II, in which a discussion is also given of various other commonly used sets of units, and of the form that familiar equations take in these other units. We shall see later how to define the volt and the weber, as well as the coulomb.

As far as charge is concerned, with our knowledge that all matter is composed of electrons, protons, neutrons, and other such elementary particles, all of which prove experimentally to carry charge equal to an integral multiple of the electronic charge e, given by

$$e = 1.60 \times 10^{-19} \text{ coulomb}, \tag{1.2}$$

we see that the determination of the charge on a body really resolves itself into a counting of the unbalanced elementary charges on it. This fact, that all charges are integral multiples of a fundamental unit, is still one of the unexplained puzzles of fundamental physics. It does not in any way contradict electromagnetic theory, but it is not predicted by it, and until we have a more fundamental theory that explains it, we shall not feel that we really understand electromagnetic phenomena thoroughly. Presumably its explanation will not come until we understand quantum theory more thoroughly than we do at present.

2. The Field of a Distribution of Static Point Charges.—For a number of chapters we shall be dealing with the forces on charges at rest, and shall consider only the electric field **E**, which will be assumed to be independent of time. This is the branch of the subject known as "electrostatics." It is now found that the field **E** can be determined, once the distribution of charge is known. This fact, together with (1.1) and Newton's laws of motion, furnishes a complete system of equations: knowing where the charge is, we find **E**; knowing **E**, we find the force **F**; knowing the force, we find the acceleration of the particles bearing the charge, and hence their motions. Of course, our special case of electrostatics must be that in which the total force, which may include nonelectrostatic as well as electrostatic forces, acting on a body, is zero, so that the body can stay at rest, and we can be dealing with a static problem.

The law giving the field **E** arising from a distribution of charge is very simple. We may consider the charge to be made up of a great many small, or point, charges (which on an atomic scale could be electrons and protons). We may compute the field resulting from each of these point charges. Then the first experimental law is that the total field is the vector sum of the fields arising from the various point charges. Furthermore, the field **E** of a static point charge q proves to be in the direction pointing away from q, and is equal in magnitude to $q/4\pi\epsilon_0 r^2$, where

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ farad/m}, \tag{2.1}$$

and where r is the distance from the charge, measured in meters.

This may for the present be considered an experimental law, with (2.1) as an experimentally determined coefficient, determined to give the result for **E** in volts per meter, if q is in coulombs. In Appendix II we shall see why it is that in terms of the definition of the volt and the coulomb we have this particular numerical constant. The factor 4π is included in the formula at this point, because it is more convenient to do so for later applications, as we shall see presently. We shall also understand later the units, farads per meter, which we have assigned to ϵ_0 in (2.1).

As a result of our simple law giving the field arising from a point charge, we can find easily the force between two charges q and q', at a distance r apart. The force is clearly in the direction joining them, and is equal in magnitude to

$$F = \frac{qq'}{4\pi\epsilon_0 r^2}. \tag{2.2}$$

It is a repulsion if q and q' have the same sign, an attraction if they have opposite sign. Equation (2.2) is Coulomb's law, in the form that it takes in the mks system; we give corresponding expressions in other systems of units in Appendix II. In Eq. (2.2) we have an illustration of the fact that problems in electrostatics can be handled by the ideas of action at a distance, since Coulomb's law is similar to the law of gravitation.

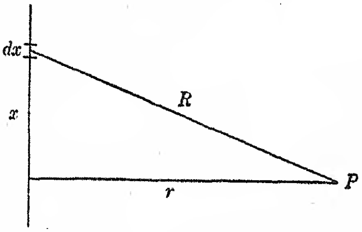


FIG. 1.—Field of a charged line.

It is such a simple matter to find the field of a point charge that we can easily sum such fields vectorially, to find the field of an arbitrary charge distribution, provided that we know where the charges are located. Simple examples of the field of distributions of particles are the fields of a uniformly charged line and plane. Let a line carry a constant charge σ per unit length. The contribution of the charge in length dx to the field at P, in Fig. 1, is along the direction of R; its component along r, which alone integrates to something different from zero, is $(\sigma\, dx/4\pi\epsilon_0 R^2)(r/R) = \sigma r\, dx/4\pi\epsilon_0 R^3$. The resultant field is

$$\frac{\sigma r}{4\pi\epsilon_0}\int_{-\infty}^{\infty} \frac{dx}{(x^2 + r^2)^{3/2}} = \frac{\sigma}{2\pi\epsilon_0 r}.$$

Similarly let a surface carry a constant charge of σ per unit area. The

contribution of the charge in the ring of radius between x and $x + dx$ in the plane, shown in Fig. 2, to the component of field along the normal, at P, is $(2\pi\sigma x\,dx/4\pi\epsilon_0 R^2)(r/R) = \sigma r x\,dx/2\epsilon_0 R^3$, and the resultant field is

$$\frac{\sigma r}{2\epsilon_0}\int_0^\infty \frac{x\,dx}{(x^2 + r^2)^{3/2}} = \frac{\sigma}{2\epsilon_0},$$

a constant independent of the distance from the plane. It should be noted, however, that the field is directed away from the plane on each side of the plane (provided σ is positive), so that there is a discon-

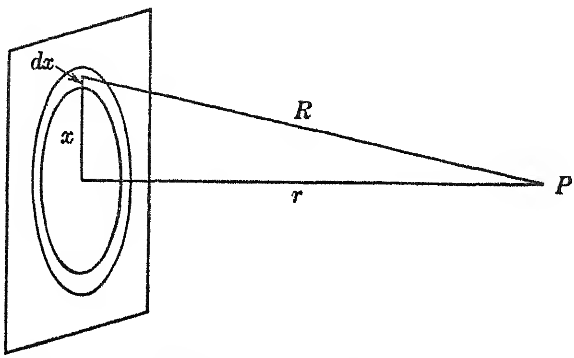


FIG. 2.—Field of a charged plane.

tinuity of σ/ϵ_0 in the normal component of E in passing through the plane. This problem of the field resulting from a distribution of charge on a plane illustrates the reason why the factor 4π was included in the statement of Coulomb's law, as in (2.2). The reason is that this factor then drops out in problems such as the field of a plane charge. If we had not included the 4π at first, it would have appeared at this point. Since more problems are similar to that of the field of a plane than to the field of a point charge, it is more convenient to handle the factor 4π as we have done.

The two examples that have just been presented show how simple it is to find fields by direct integration in some cases. We shall now go on to a more involved, but more powerful, method of finding the field of a distribution of charge, using the method of the potential. This method has two advantages over direct integration. First, it is more powerful mathematically; there are many problems to which it will provide solutions, and in which the direct integration proves to be too difficult to carry out. Secondly, there is a whole class of problems in which direct integration is of no use, for we do not know the charge

distribution to start with. This class of problems includes those involving conductors, or dielectrics. With such bodies, as we shall see, the presence of certain charges in the neighborhood induces other charges on the conductors or dielectrics, or, as we say, polarizes them. Part of our problem is to find the nature and location of this polarized charge. The method of the potential can handle this problem, whereas direct integration cannot. As a first step in taking up the potential, we consider the work done in moving a point charge from one place to another in the field, and the potential-energy function resulting from this work done.

3. The Potential.—The work that we must do in carrying a unit charge from one point to another in the field of a point charge, balancing the force exerted by the point charge, is independent of the path. To prove this, we note that the work done in going from one point to an adjacent point is the product of the force $-q/4\pi\epsilon_0 r^2$ which we exert to balance the electrostatic force, by the component of displacement dr in the direction of force, or is $-\mathbf{E}\cdot d\mathbf{s}$, where $\mathbf{E}$ is the electric field, $d\mathbf{s}$ is the vector displacement, and where the scalar product $\mathbf{E}\cdot d\mathbf{s}$ equals the product of the magnitudes of $\mathbf{E}$ and $d\mathbf{s}$, and of the cosine of the angle between them, as is discussed in Appendix I. Thus, integrating,

$$\text{Work} = -\int_1^2 \mathbf{E}\cdot d\mathbf{s} = \int_{r_1}^{r_2} -\frac{q}{4\pi\epsilon_0 r^2}\,dr = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_2} - \frac{1}{r_1}\right). \tag{3.1}$$

This work may then be written as the difference of a potential energy at the end points: if we write

$$\varphi = \frac{q}{4\pi\epsilon_0 r} \tag{3.2}$$

as the potential function at a distance r from a charge q, the work done moving a unit charge from point 1 to point 2, by (3.1), is

$$\text{Work} = \varphi_2 - \varphi_1.$$

Then, as always with problems involving potentials, we may write the force as the negative gradient of the potential:

$$\mathbf{E} = -\operatorname{grad}\varphi = -\nabla\varphi, \tag{3.3}$$

where the vector operations are discussed in Appendix I. Furthermore, using the principle of vector analysis that the curl of any gradient is zero, we have

$$\operatorname{curl}\mathbf{E} = \nabla\times\mathbf{E} = 0. \tag{3.4}$$

These relations (3.3) and (3.4) hold for the field of a point charge; by

addition, they hold for the field of any arbitrary number of point charges.

Use of this principle gives us a simplified way of computing the field of a distribution of charges. We sum the potentials of the charges, so as to get the total potential as a function of position; then we take its gradient, to get the field. This allows us to sum the potential, a scalar, rather than the field, a vector. If charges are distributed continuously in space, instead of at discrete points, we may describe them by a charge density ρ, such that $\rho\, dv$ is the charge located in volume element dv, so that ρ is in coulombs per cubic meter in the mks system. Then we may write the potential of this distribution as

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{\rho\, dv}{r}. \tag{3.5}$$

In this expression, we are finding φ as a function of x, y, and z; the integration is over $dx'\, dy'\, dz' = dv$, the coordinates of the point where the density ρ, which is really a function of x', y', z', is found. r is the distance between these two points, given by

$$r = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}.$$

The potential is expressed in volts. It will automatically satisfy (3.4), from the way in which it is built up. From φ as given in (3.5), we then find **E** by (3.3).

Surfaces φ = constant are called "equipotential surfaces." The field **E** is normal to the equipotential surfaces, as we can see either from the principle of vector analysis that the gradient of a scalar is normal to the surfaces on which the scalar is constant, or from the elementary fact that, since the work done on a charge moving from one point to another of the same equipotential surface must be zero, any displacement on that surface must be at right angles to the force vector. We may draw lines, called "lines of force," everywhere tangent to **E**; they then cut the equipotentials at right angles. These concepts, of the sort stressed particularly by Faraday, are particularly useful in electrostatic problems involving conductors. In a conductor, charges are free to move about. Thus, if conductors are present in a field, we cannot predict straightforwardly where the charge is to be found, and hence cannot find φ by (3.5), since we do not know ρ. We do know, however, that a conductor satisfies Ohm's law, which states that the current in a conductor is proportional to the electric field in it. If our problem is a static one, in which all charge is at rest

and there is no current, the field must then be zero everywhere within the conductor. The potential must then be constant, by (3.3). Thus a conductor forms an equipotential volume, its surface being an equipotential surface, so that all lines of force must cut it at right angles.

4. Electric Images.—Suppose we are given the problem of finding the potential, and hence the field, of a certain set of charges, q_1, q_2, . . . , at specified points, in the presence of certain conductors that are maintained at definite potentials. The potential must then

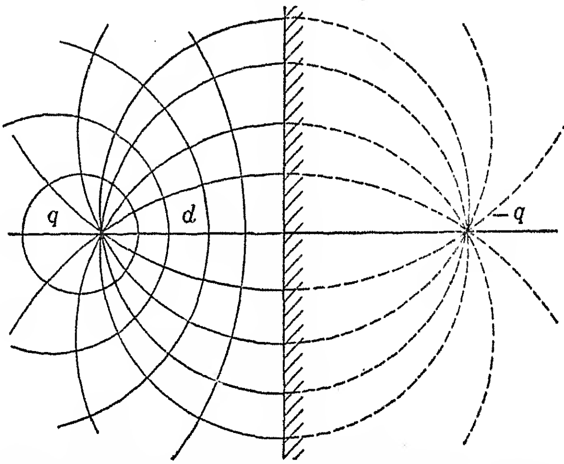


FIG. 3.—Electric image of a point charge in a conducting plane.

satisfy the following conditions: it must be equal to the required constant value on the surface of each conductor, and it must reduce to a value like (3.2) at each of the charges. These conditions can be shown to determine the potential uniquely. Sometimes by a simple device we can set up a potential, or the corresponding field, satisfying these conditions; we can then be sure we have solved our problem. Thus consider a single charge q at a distance d from a grounded conducting plane. There is a problem, shown in Fig. 3, which has a potential that reduces to the proper value where the charge is, and which gives a potential zero on the surface, by symmetry: it is the problem of two equal and opposite charges $\pm q$, at a distance $2d$ apart, the mid-plane taking the place of the surface. The correct lines of force and equipotentials are then as shown in Fig. 3, the dotted extensions being really nonexistent. The fictitious charge $-q$ behind the metallic surface is called an "electric image," from the analogy of an optical image, and this method is called the "method of images."

The method of images can also be used for the problem of a point charge and a conducting sphere. This depends on a geometrical theorem. In Fig. 4, let $oa/oc = oc/ob$. The triangles oac and obc are

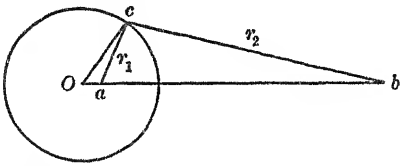


FIG. 4.—Image of a point charge in a conducting sphere.

similar, for they have the angle aoc in common, and the sides including the angle are proportional. Thus

$$\frac{oa}{oc} = \frac{oc}{ob} = \frac{r_1}{r_2}. \tag{4.1}$$

If we then place a charge q at b, and a charge $-(oc/ob)q$ at a, the potential at any point of the circle will be

$$\frac{q}{r_2} - \left(\frac{oc}{ob}\right)\frac{q}{r_1} = \frac{q}{r_2} - \frac{q}{r_2} = 0,$$

using (4.1). Thus the circle, or the sphere formed by rotating it around the axis ob, is an equipotential of zero potential. Hence the two charges, q at b and its image $-(oc/ob)q$ at a, give a field that is a solution of the problem. A somewhat more general problem can be solved by adding to the field of these two charges the field of an arbitrary point charge at the center o of the circle. This charge will give a potential that is constant over the surface of the sphere, so that the field of all three charges is consistent with the assumption that the sphere is a conductor. By adjusting the amount of charge at the center o, we can solve the problem of the potential of a point charge q at b, and a conducting sphere carrying any desired amount of charge.

Problems

1. An electron of charge $-e$ ($e = 1.60 \times 10^{-19}$ coulomb) and mass m ($m = 9.1 \times 10^{-31}$ kg) moves in a magnetic field B (in gausses) in a plane at right angles to B. Show that it moves in a circular path of radius ρ, and find the value of the product $B\rho$ in terms of the velocity of the electron.

2. An electron moves in a uniform electric field directed along the x axis, and a uniform magnetic field along the y axis. It starts from rest at the origin. Show that it moves in a cycloidal path, and find the drift velocity, or the velocity of the center of the rolling circle along the z axis. If the electric field is 10^4 volts/cm, and the magnetic field is 10^4 gausses, find the value of the drift velocity.

3. Find the angular velocity of rotation of an ion in a cyclotron. If the magnetic field is 15,000 gausses, what should be the wave length and frequency of the oscillator used to run the cyclotron, if it is accelerating protons?

4. When the velocity of a particle approaches that of light, a force at right angles to the velocity produces an acceleration transverse to the velocity given by the equation $F = m_t a$, where m_t is the transverse mass, equal to

$$\frac{m_0}{\sqrt{1 - v^2/c^2}},$$

if m_0 is the rest mass, v the velocity, c the velocity of light. Discuss the cyclotron in the relativistic range, showing that the resonant frequency must change as the ions are speeded up. If protons are to be speeded up to 500 million electron volts, find the ratio by which the oscillator frequency must change as they accelerate from rest to their maximum speed.

5. A charge e is located a distance d from an infinite conducting plane. Find how much work is required to remove the charge to infinite distance, against the attraction of its electric image.

6. Find the potential at distance r from an infinitely long straight wire carrying a charge σ per unit length.

7. Inside a hollow infinitely long grounded conducting cylinder of radius R is placed a thin charged wire, carrying a charge σ per unit length. The wire is placed parallel to the axis of the cylinder, and distant by an amount d from the axis. Find the potential at points within the cylinder. [*Hint:* Use the construction of Fig. 4 and of Eq. (4.1).]

8. Find the attraction between a charge q, and a grounded conducting sphere, as a function of the distance of q from the center of the sphere.

9. Find the attraction between a charge q, and an uncharged, insulated, conducting sphere, as a function of the distance of q from the center of the sphere. (We shall prove later that, if the sphere is uncharged, the total fictitious charge located at points o and a in Fig. 4 must be zero.)

10. Two parallel wires carry equal and opposite charges $\pm\sigma$ per unit length. Consider the intersections of the equipotentials with a plane normal to the wires, and the lines of force, which will lie in that plane. Show that the lines of force and equipotentials form two families of circles, orthogonal to each other.

11. A charge q is placed inside a hollow metal sphere of radius a at a distance r from the center. Show that the force acting on the charge q is given by

$$F = \frac{q^2 ar}{4\pi\epsilon_0(a^2 - r^2)^2}.$$

12. Find the force per unit length with which two long parallel metal cylinders, each of diameter a and separated by a distance d between centers, attract each other, if they carry a charge per unit length $\pm\sigma$, respectively.

CHAPTER II

ELECTROSTATICS

The simple concept of lines of force and of equipotential surfaces, which we introduced in the preceding chapter, becomes of real value when it forms the basis of an analytical treatment founded on differential equations. This treatment, worked out originally by Gauss, Poisson, Laplace, and others, can be used either for electrostatics, to which we apply it, or for gravitational forces, which consist of inverse-square forces, as electrostatic forces do. It was actually for the gravitational case that much of the mathematical work of the present chapter was originally developed. We shall find, however, that it forms a very natural mathematical framework for electrostatics, and that it leads us into certain equations that will later underlie the general case of electromagnetism, not merely the static problem. We develop the equations in the present chapter, and solve some electrostatic problems by means of the solution of partial differential equations in the next chapter.

1. Gauss's Theorem.—Suppose we set up a closed surface S, enclosing a volume V within which are certain charges $q_1, q_2, \ldots$. We now form a surface integral over S, as follows: We take an element da of the area of the surface. If $\mathbf{n}$ is unit vector along the outer normal, and if $\mathbf{E}$ is the value of the electric intensity at da, the surface integral is $\mathbf{E} \cdot \mathbf{n}\, da$, integrated over the surface S. Gauss's theorem is then the following:

$$\int \mathbf{E} \cdot \mathbf{n}\, da = \sum_i \frac{q_i}{\epsilon_0}. \tag{1.1}$$

That is, the surface integral of the normal component of $\mathbf{E}$ over a surface S equals $1/\epsilon_0$ times the total charge included within the surface. Let us prove this important theorem. First we prove it for a single point charge; then by addition the theorem obviously holds for any collection of point charges. For a single charge, consider an element da of surface, as shown in Fig. 5. The contribution of this element to $\mathbf{E} \cdot \mathbf{n}\, da$ is $|E|\, da \cos(\mathbf{E},\mathbf{n})$. But $da \cos(\mathbf{E},\mathbf{n})$ is the projection of da on the plane normal to $\mathbf{E}$, or is $r^2\, d\omega$, where $d\omega$ is the solid angle of the cone subtended by da. Also $|E| = q/4\pi\epsilon_0 r^2$. Thus the contri-

bution of the element to the surface integral is

$$\frac{q}{4\pi\epsilon_0 r^2} r^2\, d\omega = \frac{q}{4\pi\epsilon_0}\, d\omega.$$

Integrating over S involves integrating $d\omega$ over the complete solid angle 4π, resulting in q/ϵ_0, in agreement with (1.1). Thus, summing over an arbitrary number of point charges, Gauss's theorem is proved.

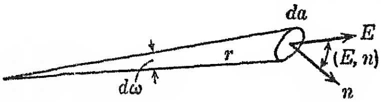


FIG. 5.—Construction for Gauss's theorem.

Examples of the application of Gauss's theorem are obvious. First consider a positive point charge, and a sphere of radius r surrounding it. By symmetry, the field points out along the radius, and depends only on r. Thus Gauss's theorem states that $|E|$ times the area, $4\pi r^2$, equals q/ϵ_0, or $|E| = q/4\pi\epsilon_0 r^2$. This is trivial, but we can equally well use Gauss's theorem to find the field of a uniformly charged line and plane, as in Chap. I, Sec. 2. For a line carrying a charge σ per unit length, let the surface S be the surface of a circular cylinder of unit length, radius r, with the line as the axis. By symmetry, the field points radially outward, and depends only on r. Thus the field is parallel to the flat ends of the cylinder, which do not contribute to the surface integral; while over the curved face the integral is $|E|$ times the area $2\pi r$. Since this equals σ/ϵ_0, we have $|E| = \sigma/2\pi\epsilon_0 r$. For the charged plane, take S as the surface of a box, as shown in Fig. 6, enclosing unit area of the surface. By symmetry, **E** will be normal to the two faces parallel to the plane. Hence only these faces will contribute to the integral. Each has unit area; thus $2|E| = \sigma/\epsilon_0$, $|E| = \sigma/2\epsilon_0$. Thus we verify the results of Chap. I, Sec. 2, by very simple methods.

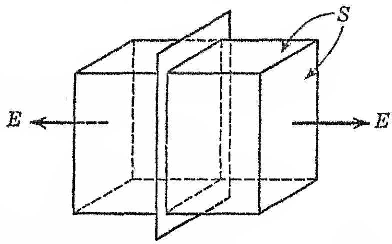


FIG. 6.—Gauss's theorem for field of charged plane.

As a final example, consider the surface of a conductor carrying a surface charge σ per unit area. We set up a surface as in Fig. 6, but now to the left of the surface, or inside the conductor, E is zero. Our volume still encloses charge σ, but now the surface integral has a contribution only from the right-hand face. Thus $|E| = \sigma/\epsilon_0$ on the right. We note that this case can be found from the earlier one by super-

posing a constant field $\sigma/2\epsilon_0$ pointing to the right on our earlier solution. This field cancels the field originally present to the left of the plane, and doubles the field originally present on the right. Such a constant field makes no contribution to the surface integral concerned in Gauss's theorem, and in fact any constant field, or more generally the field of any charge distribution outside the surface, may be added to our original solution of the problem, without contradicting Gauss's theorem. It is thus clear that Gauss's theorem by itself is not enough to determine the field completely. In our first case of the charged plane we used the additional fact of symmetry, and in the second case of the charged surface of the conductor we used the fact that the field must be zero within the conductor. In any actual case there would be information sufficient to determine the field definitely. We note an important result common to all these possible solutions for the charged plane, however: the normal component of $\mathbf{E}$ undergoes a discontinuity of σ/ϵ_0 in every case in going through a surface charge σ. Thus in the case of the charged plane the component to the right of the field changes from $-\sigma/2\epsilon_0$ to $\sigma/2\epsilon_0$, and in the case of the conducting surface it changes from zero to σ/ϵ_0. It is clear from Gauss's theorem that this relation is quite general. It has a valuable application, which will be brought out in the problems: if we know the potential and field at every point of space, including the neighborhood of conductors, we can calculate the surface-charge density on the surface of the conductors, since we know the normal component of field. We thus see, as we pointed out in the preceding chapter, that a solution for a potential that satisfies the correct boundary conditions on the surface of all conductors allows us to find the distribution of charge, though otherwise we should not know how the charge was distributed.

2. Capacity of Condensers.—A condenser consists of two conductors carrying equal and opposite charges; its capacity C is defined as the charge on one of the conductors, divided by the difference of potential between them. In a number of simple cases we can easily get the capacity by use of Gauss's theorem. Thus consider two parallel plates of area A, distance of separation d. The field $\mathbf{E}$ will be normal to the surfaces (if we neglect edge effects). If there is surface charge σ on one plate, $-\sigma$ on the other, there will be a field $E = \sigma/\epsilon_0$ between; and the difference of potential, or voltage V, between them, will be $Ed = \sigma d/\epsilon_0$. The charge Q on the plates will be σA. Thus the capacity will be

$$C = \frac{\sigma A}{\sigma d/\epsilon_0} = \frac{\epsilon_0 A}{d}.$$

From this we see the physical meaning of ϵ_0: it is the capacity of a condenser whose area is 1 sq m, with two plates 1 m apart.

Next consider a cylindrical condenser, consisting of two concentric cylinders of radii r_1, r_2 ($r_2 > r_1$), of length L. Let the inner conductor carry charge $-Q$, the outer one Q. The field between the two, by symmetry, must be radially inward, and hence by Gauss's theorem must be $-(Q/L)/2\pi\epsilon_0 r$, just as if the charge were concentrated along the axis. The potential at distance r is then the quantity whose negative gradient is the field, or is $[(Q/L)/2\pi\epsilon_0] \ln r$. Thus the difference of potential is

$$\left[\frac{Q/L}{2\pi\epsilon_0}\right](\ln r_2 - \ln r_1) = \left[\frac{Q/L}{2\pi\epsilon_0}\right]\ln\frac{r_2}{r_1}$$

and the capacity is

$$C = \frac{2\pi\epsilon_0 L}{\ln r_2/r_1}.$$

For a spherical condenser, consisting of two concentric spheres of radii r_1 and r_2 ($r_2 > r_1$), carrying charges $-Q$ and Q, respectively, the field must be radially inward, and hence by Gauss's theorem must be $-Q/4\pi\epsilon_0 r^2$, as if the charge were concentrated at the center. The potential at distance r is then $-Q/4\pi\epsilon_0 r$, and the difference of potential between conductors is $(Q/4\pi\epsilon_0)[(1/r_1) - (1/r_2)]$, so that the capacity is

$$C = \frac{4\pi\epsilon_0}{(1/r_1) - (1/r_2)}.$$

It is interesting to note that, as $r_2 \to \infty$, the capacity stays finite; this value is often referred to as the capacity of a sphere. Thus a

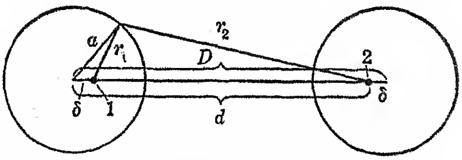


FIG. 7.—Capacity of parallel cylinder.

sphere 1 m in radius has a capacity $4\pi\epsilon_0$ with respect to an infinitely distant sphere.

As a final example of capacity we take a somewhat more complicated case, two parallel cylinders, each of radius a, length L, at distance D apart between their centers, where we assume the length to be so great that we can neglect end effects. We shall first show that equal and opposite charges $\pm Q$ distributed uniformly over the two lines 1 and 2 in Fig. 7, where $\delta/a = a/d$, will result in the two cylin-

drical conductors being equipotentials. To prove this, we use the geometrical theorem of Eq. (4.1), Chap. I, though for quite a different purpose. From it we see that $\delta/a = a/d = r_1/r_2$. A line charge $+Q$ at 1, and $-Q$ at 2, will have a potential equal to

$$\frac{Q/L}{2\pi\epsilon_0}(\ln r_1 - \ln r_2) = \frac{Q/L}{2\pi\epsilon_0}\ln\frac{r_1}{r_2}$$
$$= \frac{Q/L}{2\pi\epsilon_0}\ln\frac{a}{d}$$

at every point of the left-hand cylinder, and the negative of this at every point of the right-hand cylinder. Thus the cylinders are equipotentials, and can be replaced by conductors. By Gauss's theorem, the total charge on each cylinder is $\pm Q$; for the surface integral of the normal component of **E** over the surface of the cylinder must be $1/\epsilon_0$ times the total charge inside, which is $\pm Q$ for the line charges, and hence for the cylinders that produce equal fields at external points. Thus the difference of potential between the cylinders is $[(Q/L)/\pi\epsilon_0] \ln a/d$, so that the capacity is

$$C = \frac{\pi\epsilon_0 L}{\ln a/d}.$$

This can be put in terms of D, the distance between centers, rather than d. We have

$$\frac{D}{a} = \frac{d}{a} + \frac{\delta}{a} = \frac{d}{a} + \frac{a}{d} = e^{\pi\epsilon_0 L/C} + e^{-\pi\epsilon_0 L/C} = 2\cosh\frac{\pi\epsilon_0 L}{C},$$

or

$$C = \frac{\pi\epsilon_0 L}{\cosh^{-1} D/2a}. \tag{2.1}$$

3. Poisson's Equation and Laplace's Equation.—Suppose that within a volume V we have, not a discrete set of point charges, but a continuous distribution of volume charge, such as we should have for instance from an electronic space charge, if our scale of measurement is large enough compared with the interelectronic distance. We can then define a charge density ρ as the charge per unit volume (in coulombs per cubic meter in the mks system). Then Gauss's theorem is replaced by

$$\int_S \mathbf{E}\cdot\mathbf{n}\,da = \frac{1}{\epsilon_0}\int_V \rho\,dv, \tag{3.1}$$

where the volume integral is over the volume V enclosed by the surface

S. This must hold for any arbitrary surface and volume. Now we may apply the divergence theorem of vector analysis, discussed in Appendix I. This states that, if $\mathbf{F}$ is any vector function of position, and div $\mathbf{F}$, or $\nabla \cdot \mathbf{F}$, is its divergence, then

$$\int_S \mathbf{F} \cdot \mathbf{n}\, da = \int_V \operatorname{div} \mathbf{F}\, dv. \tag{3.2}$$

Applying to (3.1), we have

$$\frac{1}{\epsilon_0} \int_V \rho\, dv = \int_V \operatorname{div} \mathbf{E}\, dv,$$

where the equation holds for any arbitrary volume V. This cannot be the case unless the integrands are equal; that is, unless

$$\operatorname{div} \mathbf{E} = \frac{\rho}{\epsilon_0}. \tag{3.3}$$

This is one of the fundamental equations of electromagnetic theory. We now combine with Eq. (3.3) of Chap. I, $\mathbf{E} = -\operatorname{grad} \varphi$, where φ is the potential. We then have

$$\nabla^2 \varphi = -\frac{\rho}{\epsilon_0}. \tag{3.4}$$

This is Poisson's equation. In a region where there is no volume charge, so that $\rho = 0$, Poisson's equation reduces to

$$\nabla^2 \varphi = 0, \tag{3.5}$$

which is Laplace's equation, and is one of the most important equations in mathematical physics. It is as a result of its appearance in this equation that the operator ∇^2 is called the "Laplacian."

As an illustration of the variety of problems in which Laplace's equation is encountered, we may mention static elasticity; elastic vibrations are governed by a wave equation, and when the time variation in the wave equation is set equal to zero, Laplace's equation results. Another illustration is the steady-state flow of heat, where the temperature satisfies Laplace's equation. Diffusion, and the steady flow of electricity, are two other examples. Many more complicated equations, such as the wave equations of wave mechanics, contain the Laplacian operator, and are solved by analogy with Laplace's equation, so that a knowledge of the solutions of Laplace's equation proves to underlie a great deal of mathematical physics.

4. Green's Theorem, and the Solution of Poisson's Equation in an Unbounded Region.—A partial differential equation, like Poisson's

or Laplace's equation, has a great variety of solutions. In the first place, Poisson's equation, being a linear inhomogeneous differential equation (that is, containing the Laplacian, involving first powers of derivatives of φ, and containing the term $-\rho/\epsilon_0$ involving terms independent of φ), has a general solution that can be formed as follows: we find a particular solution, or particular integral, of the inhomogeneous equation. We also find a general solution of the related homogeneous equation, formed by setting the term independent of φ equal to zero [that is, in this case, of Laplace's equation (3.5)]. This homogeneous equation is called the "auxiliary equation," and a general solution of it is called the "complementary function." Then the sum of the particular integral, and of the complementary function, is in the first place a solution of the inhomogeneous equation, as we can see at once by substituting it in the equation. Furthermore, if the complementary function has sufficient arbitrariness in it, we can use it to satisfy arbitrary boundary conditions. Thus the sum of complementary function and particular integral is a complete solution of the inhomogeneous equation. In the next chapter we shall take up some methods for solving Laplace's equation, and satisfying arbitrary boundary conditions with it. We shall find that, to do this, we require an infinite number of arbitrary constants, or an arbitrary function, since we are dealing with a partial differential equation. In the present chapter we shall demonstrate a simple method, called "Green's method," for finding a particular integral of (3.4). By combining the two, we shall then be in position to get a complete solution of Poisson's equation.

The first step in deriving Green's solution of Poisson's equation is to prove Green's theorem, an important theorem in vector analysis, discussed in Appendix I. We prove it from the divergence theorem, (3.2), by setting $\mathbf{F} = \varphi \operatorname{grad} \psi$, where φ and ψ are scalars. Then (3.2) becomes

$$\int_S \varphi \operatorname{grad} \psi \cdot \mathbf{n}\, da = \int_V (\varphi \nabla^2 \psi + \operatorname{grad} \varphi \cdot \operatorname{grad} \psi)\, dv. \qquad (4.1)$$

This is one form of Green's theorem. To get the more familiar form, we next write the same expression with φ and ψ interchanged, and subtract, obtaining

$$\int_S (\varphi \operatorname{grad} \psi \cdot \mathbf{n} - \psi \operatorname{grad} \varphi \cdot \mathbf{n})\, da = \int_V (\varphi \nabla^2 \psi - \psi \nabla^2 \varphi)\, dv. \quad (4.2)$$

We can now use this theorem to obtain Green's solution of Poisson's equation. Let us apply the theorem to the whole of space outside

a very small sphere of radius R surrounding a point P. Let $\psi = 1/r$, where r is the distance from P to the point x, y, z. Then $\nabla^2\psi = 0$ everywhere outside the sphere of radius R; for ψ is the potential of a charge located at P, and hence satisfies Laplace's equation everywhere except where the charge is located, or at P. The right side of (4.2) is then

$$-\int \frac{\nabla^2\varphi}{r}\,dv,$$

integrated over all space except the small sphere. For the left side, $\operatorname{grad}\psi \cdot \mathbf{n} = -d(1/r)/dr = 1/r^2$. Thus the left side is

$$\int \frac{\varphi}{r^2}\,da + \int \frac{1}{r}\frac{\partial\varphi}{\partial r}\,da.$$

On the surface of the sphere, $r = R$, $\int(\varphi/r^2)\,da$ is then

$$\left(\frac{1}{R^2}\right)\bar{\varphi}\int da = 4\pi\bar{\varphi},$$

where $\bar{\varphi}$ is the average value of φ over the sphere, or is approximately the value of φ at P. The second term is $(1/R)(\overline{\partial\varphi/\partial r})(4\pi R^2)$, which goes to zero as R approaches zero. Thus in the limit as R goes to zero, we have

$$4\pi\varphi = -\int \frac{\nabla^2\varphi}{r}\,dv. \tag{4.3}$$

This is a mathematical theorem holding for any function φ, where the volume integral on the right side is extended over all space except an infinitesimal sphere surrounding the point P, where φ is the value of φ at P, and where r is the distance from P to x, y, z. Now let us combine this with Poisson's equation (3.4). Then, for the potential φ of the charge ρ, we have

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{\rho}{r}\,dv. \tag{4.4}$$

This is Green's solution of Poisson's equation. It is identical with Eq. (3.5) of Chap. I, which we have already obtained; it is simply a more elegant method of deriving that result. We shall find this more elegant method to be useful, in the next chapter, in finding the potential in a bounded region of space, rather than an unbounded region as we have here.

5. Direct Solution of Poisson's Equation.—Green's method forms one way of solving Poisson's equation. However, sometimes ρ has a

sufficiently simple form so that we can solve it directly, regarding it as a problem in differential equations. Thus, for instance, suppose ρ is a function of one variable only, say x, independent of y and z. Then the potential may also be chosen to be independent of y and z. In this case Poisson's equation becomes an ordinary differential equation:

$$\frac{d^2\varphi}{dx^2} = -\frac{\rho(x)}{\epsilon_0}.$$

This can be solved as an ordinary differential equation for φ as a function of x, merely by integrating twice with respect to x. As a very simple case, let ρ be a constant. Then we have simply

$$\varphi = \varphi_0 - E_0 x - \frac{1}{2}\frac{\rho}{\epsilon_0}x^2. \tag{5.1}$$

Such solutions of Poisson's equation are useful in space-charge problems. In studying space-charge limited emission from cathodes of various shapes, we use Poisson's equation, expressed in variables appropriate to the cathode geometry, to find the potential from the charge distribution. This is combined with the dynamical equations of motion of the electrons, Newton's law giving the acceleration in terms of the force derived from the potential, and in this way we get a complete set of equations to determine both charge density and potential.

We must remember, as has been mentioned earlier, that the solution, as found for instance in (5.1), is by no means a general solution of Poisson's equation, even though it contains the two arbitrary constants, φ_0 and E_0. For it is clear that we can add to it any solution of Laplace's equation, acting as the complementary function. It is true that our function $\varphi_0 - E_0 x$ is a solution of Laplace's equation, but it is far from a general solution, since it is a function of x only. We shall examine general solutions of Laplace's equation in the next chapter, and shall show that a great variety of such solutions exists. We shall show that, by using a general solution of Laplace's equation, we can satisfy general boundary conditions: we can find a solution such that φ reduces to prescribed values over the boundary of the region in which we can satisfy Poisson's or Laplace's equation.

Problems

1. Given a spherical distribution of charge, in which the density is a function of r. Prove that the field at any point is what would be obtained by imagining a

sphere drawn through the point, with its center at the origin, all the charge within the sphere concentrated at the center, and all the charge outside removed. Apply this result to the gravitational case, showing that the earth acts on bodies at its surface as if its mass were concentrated at the center.

2. Given a sphere filled with charge of constant density. Prove that, at points within the sphere, the field is directly proportional to the distance from the center.

3. Find the surface density induced by a charge on a plane conductor. Show by direct integration that the total induced surface charge equals the inducing charge in magnitude.

4. For a certain spherical distribution of charge, the potential is given by $\frac{-qe^{-ar}}{4\pi\epsilon_0 r}$, where q, a are constants. Find a distribution of charge that will produce this potential, finding charge density as function of r, and the charge contained between r and $r + dr$. Consider whether a point charge at the origin is also required to produce the potential. The resulting charge distribution represents roughly the charge distribution within an atom.

5. There are certain charges and conductors in an electrostatic field, whose potential is φ. Show that the surface density of charge on the surface of a conductor is $-\frac{1}{\epsilon_0}\frac{\partial\varphi}{\partial n}$, where n is the normal pointing out of the conductor.

6. Prove that the potential cannot have a minimum in an uncharged region of space. Prove therefore that a point charge cannot be in stable equilibrium under the action of electrostatic forces in an uncharged region.

7. A certain vacuum tube contains a cylindrical cathode of radius r_1. Surrounding it is a space-charge sheath, of constant density $-\rho$ (where ρ is positive), extending to a larger radius r_2. The anode is a cylinder of still larger radius r_3. If the cathode is at potential φ_1, the anode at potential φ_3, find the potential as a function of r, both inside and outside the space charge.

8. Discuss the charge distribution giving rise to equipotentials given by the real part of $-\frac{q}{2\pi\epsilon_0}\ln[\sin\pi(x + jy)] = \text{constant}$. What is the form of these equipotentials for large values of y? Sketch the lines of force and the equipotentials.

CHAPTER III

SOLUTIONS OF LAPLACE'S EQUATION

A partial differential equation, such as Laplace's equation, has a great variety of solutions. The type of solution to be used depends largely on the type of boundary condition that must be satisfied. One method of solution is an extension of Green's method; we meet it later in this chapter. Another method is called "separation of variables." This involves finding solutions that are the product of a separate function of each of the three coordinates. It is a common method for solution of similar equations, and is in fact the most powerful method available for the purpose. Different solutions can be found, depending on whether we work in rectangular, polar, or other coordinate systems. We start with rectangular coordinates, for a simple illustration of the method.

1. Solution of Laplace's Equation in Rectangular Coordinates by Separation of Variables.—Working in rectangular coordinates x, y, z, let us try to find a solution of $\nabla^2\varphi = 0$ in which φ is a product of a function X of x, a function Y of y, and a function Z of z (this is certainly not the most general solution, but we shall find that solutions of this type exist). Then Laplace's equation becomes

$$YZ\frac{d^2X}{dx^2} + ZX\frac{d^2Y}{dy^2} + XY\frac{d^2Z}{dz^2} = 0.$$

We divide by XYZ, obtaining

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} = 0.$$

In this equation, the first term is a function of x only, the second a function of y only, the third a function of z only. It is clearly impossible that the sum of these should be zero independent of x, y, z, unless each term separately is a constant, and the constants add to zero. That is, we must have

$$\frac{1}{X}\frac{d^2X}{dx^2} = a^2, \quad \frac{1}{Y}\frac{d^2Y}{dy^2} = b^2, \quad \frac{1}{Z}\frac{d^2Z}{dz^2} = c^2, \quad a^2 + b^2 + c^2 = 0. \qquad (1.1)$$

From these we then have three ordinary differential equations,

$$\frac{d^2X}{dx^2} = a^2X, \qquad \frac{d^2Y}{dy^2} = b^2Y, \qquad \frac{d^2Z}{dz^2} = c^2Z,$$

whose solutions are

$$\begin{aligned} X &= A_1e^{ax} + A_2e^{-ax}, \\ Y &= B_1e^{by} + B_2e^{-by}, \\ Z &= C_1e^{cz} + C_2e^{-cz}, \end{aligned}$$

where as a result of (1.1), some of the constants a, b, c must be real and some imaginary. Thus some of the functions X, Y, Z vary exponentially with the arguments, the others sinusoidally. The product XYZ is now a particular solution of Laplace's equation.

Any linear combination of such particular solutions is also a solution. Thus as a general solution we have

$$\varphi = \Sigma(A_1e^{ax} + A_2e^{-ax})(B_1e^{by} + B_2e^{-by})(C_1e^{cz} + C_2e^{-cz})$$

where the summation is over an infinite number of terms, each with separate constants A_1, A_2, B_1, B_2, C_1, C_2, a, b, c, subject only to the relation $a^2 + b^2 + c^2 = 0$. Thus the general solution has an infinite number of arbitrary constants. This is characteristic of partial differential equations. The constants are to be chosen so as to fit the boundary conditions. We shall use rectangular coordinates if the boundary conditions are imposed over the surface of a rectangular region. For instance, suppose we have two planes, $x = 0$ and $x = d$, on which the potentials are given as functions of y, independent of z; we might have a composite electrode on each of these planes, different strips being maintained at different potentials by batteries. Suppose we wish to find the potential in the region between the planes. Clearly we can let φ be independent of z, so that $c = 0$, and $a^2 = -b^2$. Then we may write

$$\varphi = \Sigma(A_1e^{ax} + A_2e^{-ax})(B_1e^{jay} + B_2e^{-jay}).$$

where $j = \sqrt{-1}$. When $x = 0$, this becomes

$$\Sigma(A_1 + A_2)(B_1e^{jay} + B_2e^{-jay}),$$

and when $x = d$ it is

$$\Sigma(A_1e^{ad} + A_2e^{-ad})(B_1e^{jay} + B_2e^{-jay}).$$

Now a sum of an infinite number of terms of the form ΣDe^{jay} is one way of writing a Fourier series, which we discuss in Appendix III. Thus at $x = 0$ the potential must be given by a Fourier series with

coefficients $(A_1 + A_2)B_1$ for the term in e^{iay}, and at $x = d$ it is given by another Fourier series with coefficients $(A_1e^{ad} + A_2e^{-ad})B_1$. By the theory of Fourier series we can find these coefficients, and hence the quantities A_1B_1, A_2B_1 (and similarly A_1B_2, A_2B_2). Thus we can get the infinite set of arbitrary constants, from the boundary conditions.

The method we have used above, for satisfying boundary conditions in Laplace's equation, can be used also in discussing Poisson's equation. Thus suppose we have the same problem as above, two parallel planes on which the potentials are to be prescribed functions of y, but that between the planes we have certain charge distributions that are assumed given. To keep the simplicity of our problem resulting from the independence of z, these charge distributions as well should be independent of z. We now know, from the two preceding chapters, how to find a particular solution of Poisson's equation for the potential resulting from these charge distributions. In particular, for this problem, the charge can be considered as made up of uniform charges along lines parallel to the z axis, and we have already seen that such a charged line has a logarithmic potential, so that all we need to do is to sum such logarithmic potentials arising from all linear charges. The resulting potential, however, will not have the correct values along the two planes. We then set up the complementary function, a solution of Laplace's equation, built up as in the solution we have worked out in this section, and reducing to boundary values on the surfaces which, added to the potential already found from our particular solution of Poisson's equation, will give the desired values. This is a problem of the type we have already considered, so that we see that the determination of the complementary function for solving Poisson's equation is a problem exactly analogous to solving Laplace's equation directly.

2. Laplace's Equation in Spherical Coordinates.—The method of separation of variables, which we have applied in rectangular coordinates in the preceding section, can be applied as well in a number of other coordinate systems, notably in cylindrical and spherical coordinates, and as a less familiar case in ellipsoidal coordinates. The case of spherical coordinates is the more important in practice, for it is used for the field of point charges, dipoles, and other charge distributions concentrated near a point. In spherical polar coordinates r, θ, φ, Laplace's equation for the potential ψ (we use this symbol, so as not to confuse it with the coordinate φ) is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\varphi^2} = 0,$$

as we see in Appendix IV, in which we discuss vector operations in curvilinear coordinates. We assume that ψ is a product of three functions, one a function of r, one a function of θ, one a function of φ. That is, we assume $\psi = R(r)\Theta(\theta)\Phi(\varphi)$. Inserting this function, and dividing by the product $R\Theta\Phi$, we have

$$\frac{1}{R}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{1}{\Theta}\frac{1}{r^2 \sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \frac{1}{\Phi}\frac{1}{r^2\sin^2\theta}\frac{d^2\Phi}{d\varphi^2} = 0.$$

If we multiply by $r^2\sin^2\theta$, the first term and the second will depend on r and θ only, the last on φ only. This is impossible unless each of these is a constant, and the constants add to zero. If the last term is $-m^2$, we shall have

$$\frac{d^2\Phi}{d\varphi^2} + m^2\Phi = 0 \tag{2.1}$$

$$\frac{1}{R}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{1}{\Theta}\frac{1}{r^2\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{m^2}{r^2\sin^2\theta} = 0. \tag{2.2}$$

If we multiply by r^2, the first term will depend only on r, the second and third only on θ. Thus again each of these must be a constant. Let the first term be $l(l+1)$. Then

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) - \frac{l(l+1)}{r^2}R = 0, \tag{2.3}$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left[l(l+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0. \tag{2.4}$$

We have now separated the variables, in the sense that we have three ordinary differential equations, (2.3), (2.4), (2.1), for the functions R, Θ, Φ.

3. Spherical Harmonics.—The solution of these equations is not difficult. For (2.3) we may try the assumption $R = r^n$. Then we find that we have a solution if $n(n+1) = l(l+1)$, whose solutions are $n = l$, or $n = -(l+1)$. Thus

$$R = ar^l + \frac{b}{r^{l+1}},$$

where a, b are constants of integration. Equation (2.4) is Legendre's equation. It can be solved by making the substitution

$$\Theta = \sin^m\theta(A_0 + A_1\cos\theta + A_2\cos^2\theta + \cdots).$$

We then find, on substituting in (2.4), that the following relations

must exist between the A's:

$$\begin{aligned} A_0[l(l+1) - m(m+1)] + 2A_2 &= 0 \\ A_1[l(l+1) - (m+1)(m+2)] + 2 \cdot 3A_3 &= 0 \\ A_2[l(l+1) - (m+2)(m+3)] + 3 \cdot 4A_4 &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & \end{aligned}$$

These allow us to solve for all the A's in terms of A_0 and A_1, which are constants of integration. In this way we find

$$\begin{aligned} \Theta = A_0 \sin^m \theta \Bigg\{ 1 &- \frac{[l(l+1) - m(m+1)]}{2!} \cos^2 \theta \\ &+ \frac{[l(l+1) - m(m+1)][l(l+1) - (m+2)(m+3)]}{4!} \cos^4 \theta - \cdots \Bigg\} \\ + A_1 \sin^m \theta \Bigg\{ \cos \theta &- \frac{[l(l+1) - (m+1)(m+2)]}{3!} \cos^3 \theta \\ &+ \frac{[l(l+1) - (m+1)(m+2)][l(l+1) - (m+3)(m+4)]}{5!} \cos^5 \theta \\ &- \cdots \Bigg\}. \end{aligned} \tag{3.1}$$

Finally the solution of (2.1) is

$$\Phi = C \sin m\varphi + D \cos m\varphi, \tag{3.2}$$

where C, D are constants of integration.

We now find that m and l must be chosen to be integers to satisfy certain conditions. Unless m is an integer, an increase of φ by 2π, which brings us back to the same point of space, would lead to a different value of Φ; thus m is an integer, provided that we are solving Laplace's equation in a region that includes all values of φ. Next consider (3.1). If the series do not break off, they can be shown to diverge when $\cos \theta = \pm 1$, or along the axis of coordinates. To show that this is plausible, we find the ratio of the term in $\cos^p \theta$ to that in $\cos^{p-2} \theta$. This ratio is

$$-\frac{[l(l+1) - (m+p-2)(m+p-1)]}{p(p-1)} \cos^2 \theta,$$

which approaches $\cos^2 \theta$ in the limit where p is very large. Thus for $\cos \theta = \pm 1$, this ratio approaches $+1$, so that the terms of the series oscillate without decreasing in magnitude. Under these circumstances it is possible, though not necessary, for a series to diverge. It is also possible for the series to converge, but a closer examination than we shall give shows in the present case that the series actually

diverges, and the series would become infinite for $\cos\theta = \pm 1$. If we are solving Laplace's equation in a region including the axis, the potential certainly cannot become infinite along the axis. Thus the series cannot be allowed to diverge. This can be prevented only if l is an integer; for then one of the series (3.1) breaks off to form a polynomial, which is finite for all values of θ. The corresponding polynomials are called "associated Legendre functions," and are commonly denoted by $P_l^m(\cos\theta)$, provided that the constant factor multiplying them is properly chosen. These constant factors, and additional properties of the associated Legendre functions, are taken up in Appendix V. The product of $\sin m\varphi$ or $\cos m\varphi$ by $P_l^m(\cos\theta)$ is called a "spherical harmonic." We then find as our general solution

$$\psi = \left(ar^l + \frac{b}{r^{l+1}}\right) P_l^m(\cos\theta)(C \sin m\varphi + D \cos m\varphi). \qquad (3.3)$$

Here, as with the rectangular case, we can add an infinite number of such terms to get a general solution, and the constants can be chosen so as to fit boundary conditions. For instance, we may solve the problem of the potential between two concentric spheres of radii r_1, r_2, where the potential is given by specified functions of θ, φ on each of the spheres. The method of satisfying these boundary conditions is taken up in Appendix V.

4. Simple Solutions of Laplace's Equation in Spherical Coordinates. We have derived a general solution (3.3) of Laplace's equation. This solution is of great value in complicated problems that can be expressed in terms of spherical coordinates. However, in certain simple cases the solution reduces to very simple forms, which are of great importance. Thus let us ask for a solution corresponding to $l = 0$, $m = 0$. In this case, by (3.1) and (3.2), we see that the functions of θ and φ are constants, so that ψ reduces to $a + b/r$. The first term is a constant of integration, the second the potential of a charge located at the origin. We can determine the constant b if we know the charge, by using Gauss's theorem, or the fact that the potential of a charge q at the origin is $q/4\pi\epsilon_0 r$, so that $b = q/4\pi\epsilon_0$.

The next simplest solution is for $l = 1$, $m = 0$. In this case the first sum of (3.1) fails to break off, and diverges, so that we must set its coefficient equal to zero in our solution; but the second breaks off after the first term, and is a constant times $\cos\theta$. Thus for this case we have

$$\psi = \left(ar + \frac{b}{r^2}\right)\cos\theta. \qquad (4.1)$$

The first term, proportional to $r\cos\theta$, is simply proportional to z, the coordinate along the axis of the spherical coordinates; thus its field is a constant along the z axis, which is surely a solution of Laplace's equation. The next term, proportional to $\cos\theta/r^2$, is particularly interesting: it is the potential of a dipole, which we shall consider in the next section. Before proceeding to it, we note that (4.1) can be used to solve an interesting problem: the potential of a grounded conducting sphere in a uniform external field. Let the external field be along the z axis, and let its magnitude be E. Then at large distances from the sphere we must have $\psi = -Ez = -Er\cos\theta$, or we must have $a = -E$ in (4.1). Next let the radius of the sphere be R, so that the potential must be zero on the sphere. We can arrange to have the potential of (4.1) satisfy this condition, by making

$$aR + \frac{b}{R^2} = 0, \qquad b = ER^3.$$

Thus (4.1) reduces to

$$\psi = \left(-Er + \frac{ER^3}{r^2}\right)\cos\theta, \tag{4.2}$$

which is a solution of Laplace's equation, reduces to zero on the surface of the conducting sphere, and reduces to the correct value at infinite distance.

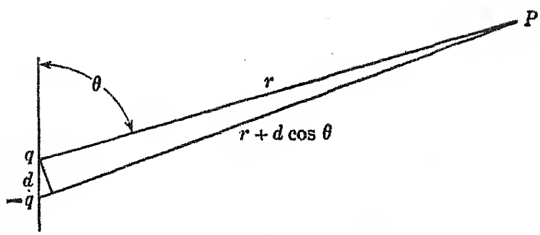


FIG. 8.—Potential of a dipole.

5. The Dipole and the Double Layer.—A dipole by definition consists of two equal and opposite charges $\pm q$ separated by a distance d. The potential at point P is then

$$\frac{q}{4\pi\epsilon_0}\left(\frac{1}{r} - \frac{1}{r + d\cos\theta}\right),$$

where r, θ are as shown in Fig. 8. If d is small, this is approximately

$$\frac{q}{4\pi\epsilon_0}\frac{d\cos\theta}{r^2} = -\frac{qd}{4\pi\epsilon_0}\left[\operatorname{grad}\left(\frac{1}{r}\right)\cdot\mathbf{n}\right],$$

where $\mathbf{n}$ is the unit vector along the axis of the dipole. The product qd is called the "dipole moment," and is denoted by m. Thus the potential of the dipole is

$$\psi = \frac{m}{4\pi\epsilon_0} \frac{\cos\theta}{r^2}. \tag{5.1}$$

In speaking of dipoles, it is customary to assume that q is infinitely large, d infinitesimally small, in such a way that the dipole moment m is finite, and (5.1) forms the exact potential. By comparison with (4.1) we see that the second term of that expression is the potential of a dipole of moment $4\pi\epsilon_0 b$, located at the origin.

By analogy with this treatment of the dipole and its potential, one can set up solutions of Laplace's equations similar to (4.1), but corresponding to larger values of l than unity. One can then interpret these solutions as being the potentials of more complicated distributions of charge at the origin. In this way one arrives at the definitions of quadripoles, and a whole series of higher multipoles, corresponding to successively larger values of l. We consider these higher multipoles in Appendix VI. They are of particular importance in studying the external electric fields (and, as we shall see later, of magnetic fields) of atoms and molecules on the one hand, and of atomic nuclei and their elementary constituents on the other. We often are interested in their fields at a large distance compared with their dimensions, and then they act to a good approximation like multipoles. Thus, for instance, a diatomic molecule like HCl, one of whose atoms (H) tends to acquire a positive charge, the other (Cl) a negative charge, has a field at a distance that consists of a dipole as its leading term. On the other hand, a molecule like N_2, which is symmetrical, has no dipole moment. We notice from (3.3) that the higher l is, the more rapidly the potential falls off with distance, varying as $r^{-(l+1)}$. Thus, since the higher multipoles correspond to higher values of l, their fields fall off more rapidly, and at large distances the field of the multipole of lowest l value predominates. For nuclei, the leading term in the field at large distances, aside from the electrostatic potential varying with $1/r$ resulting from the nuclear charge, is generally a magnetic rather than an electrical term.

The reader might reasonably be surprised that in our present discussion we started by trying to solve Laplace's equation, which is the special case of Poisson's equation which we meet in a region in which there are no charges, and end up by finding the potential resulting from a point charge, dipole, or higher multipole. How, he might ask, did charges get into the theory? The answer is that our

spherical coordinate system introduces a singularity at the origin, and our final solutions, although they satisfy Laplace's equation perfectly properly everywhere except at the origin, fail to satisfy it at the origin. The term in (3.3) in r^l has no singularity at the origin, and if we know that there is no charge distribution of any sort at the origin, that term alone must be used in setting up a solution of Laplace's equation. On the other hand, the terms in r^l become infinite at infinite distance from the origin. Closer examination of them shows that they must be connected with the existence of infinite amounts of charge at infinite distance. Thus, to take a very simple case, suppose we have the potential $r \cos \theta$, corresponding to the first term of (4.1). This, as we have seen, is the potential of a uniform field, such as would be set up within an infinite parallel plate condenser. To set it up, we should then have to have equal and opposite charge distributions over two infinitely great condenser plates at infinite distance, and each would carry infinite charge. If, then, by some condition of a problem, we know that there is no charge at infinite distance, we must not use the terms in r^l in the solution (3.3) of Laplace's equation. Of course, if there is no charge at infinity, so that we do not use the terms in r^l, and no charge at the origin, so that we do not use the terms in $r^{-(l+1)}$, and no charge anywhere between, as is implied by the fact that we are trying to solve Laplace's equation, then there will be no field, the potential will be constant everywhere, and there is no problem to be solved.

Returning to the discussion of dipoles, we have seen that (5.1) represents the potential of a dipole of moment m. Similarly the potential found in (4.2), for the problem of a grounded conducting sphere in a uniform external field, is the potential of the superposition of a uniform field, and of a dipole of moment $4\pi\epsilon_0 R^3 E$. We thus see that a conducting sphere of radius R acquires an induced dipole moment of this amount in an external field. When an object thus acquires a dipole moment, it is said to be "polarized," and if the moment is proportional to the field, the ratio is defined as the polarizability α. Thus we have

$$m = \alpha E. \tag{5.2}$$

As we have just seen, the polarizability of a conducting sphere of radius R is $4\pi\epsilon_0 R^3$.

Often we have occasion to consider, not a single dipole, but a so-called "double layer," a surface distribution of dipoles. Thus let two surfaces a distance d apart carry surface charges $\pm\sigma$ per unit

area. Then we may say that the dipole moment per unit area is σd. We note that this double layer is equivalent to a parallel plate condenser, and that the potential difference between the plates is $\sigma d/\epsilon_0$, or the dipole moment per unit area divided by ϵ_0. The field is zero outside the double layer. We note that the normal component of $\mathbf{E}$ is discontinuous by amount σ/ϵ_0 at a charged surface (as seen in Sec. 1, Chap. II), and the potential is discontinuous by an amount (dipole moment per unit area) divided by ϵ_0 at a double layer.

6. Green's Solution for a Bounded Region.—So far in this chapter we have been considering the direct solution of Laplace's equation, by separation of variables, in rectangular or spherical coordinates. We now consider the other important solution, using Green's method. We proceed as in Chap. II, Sec. 4, only we apply Green's theorem to the volume V between the small sphere of radius R surrounding the point P, and a larger surface S'. Then, substituting $\psi = 1/r$ in Green's theorem, (4.2) of Chap. II, and proceeding as in the derivation of (4.3) of Chap. II, we have in addition a surface integral over S'. We then have in general

$$4\pi\varphi = -\int_V \frac{\nabla^2\varphi}{r}\,dv - \int_{S'}\left(\varphi \text{ grad } \frac{1}{r}\cdot \mathbf{n} - \frac{1}{r}\text{ grad } \varphi\cdot\mathbf{n}\right)da.$$

This, like (4.3) of Chap. II, is a general theorem holding for any function φ. If in particular $\nabla^2\varphi = -\rho/\epsilon_0$, we have

$$\varphi = \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho}{r}\,dv - \int_{S'}\left[\frac{\varphi}{4\pi}\left(\text{grad }\frac{1}{r}\right)\cdot\mathbf{n} - \frac{1}{4\pi r}\,(\text{grad }\varphi\cdot\mathbf{n})\right]da. \quad (6.1)$$

This expresses φ at any point within V as the sum of two terms: first, the volume integral, representing the potential of the charges within V; secondly, the surface integral, which can be computed if the values of φ and its normal derivative are known at all points of the surface S'. Thus this theorem is in proper form to use in a case in which the behavior of the potential is known around the boundary of a region. Using the results of the preceding section, we can give a physical interpretation to the terms of the surface integral. Suppose we tried to set up such a distribution of surface-charge density, and such a double layer, on S', that the field and potential within S' would have the values actually present in the problem, but so that the field and potential outside S' would be zero. The normal component of field on the inner side of S' would be $-(\text{grad }\varphi)\cdot\mathbf{n}$, so that the discontinuity in field at the surface is $(\text{grad }\varphi)\cdot\mathbf{n}$. This discontinuity of

field would be produced by a surface charge of density $\epsilon_0(\text{grad } \varphi) \cdot \mathbf{n}$. The contribution of this charge to the potential would then be

$$\frac{\int \epsilon_0(\text{grad } \varphi) \cdot \mathbf{n} \, da}{4\pi\epsilon_0 r},$$

which is the second term of the surface integral in (6.1). Similarly the potential within S' is φ, so that the discontinuity is $-\varphi$, and the corresponding dipole moment per unit area to produce this discontinuity is $-\epsilon_0\varphi$. The contribution of this to the potential is

$$\int \frac{\epsilon_0\varphi}{4\pi\epsilon_0}\left(\text{grad } \frac{1}{r}\right) \cdot \mathbf{n} \, da,$$

where the differentiation in the gradient is with respect to the coordinates of the point P. It is the negative of this if the differentiation is with respect to the coordinates x, y, z of the other end of the vector r, as in (6.1). The integral is then just the first term of the surface integral in (6.1). The distribution of surface charge and double layer over S' which we have just discussed is called "Green's distribution." We now see the significance of (6.1): the potential produced by the charge ρ within V, and Green's distribution over S', is just the potential φ that we desire within V, but is zero everywhere outside.

Problems

1. It takes several volts of energy to remove an electron from the interior of a metal to the region outside. Find how many volts, if we represent the surface layer of the metal by a double layer consisting of two parallel sheets of charge; a sheet of negative electricity, of density as if there were electrons of charge 1.60×10^{-19} coulomb, spread out uniformly with a density of one to a square 4×10^{-8} cm on a side; and inside that at a distance of 0.5×10^{-8} cm a similar sheet of positive charge.

2. Find the components of electric field resulting from a dipole, both in rectangular and in spherical coordinates.

3. A charge q is at a distance r from a polarizable dipole, whose moment is α times the field in which it is located, where α is the polarizability. Find the force of attraction between charge and dipole.

4. Referring to Prob. 9, Chap. I, show that the force between a charge q and an uncharged conducting sphere can be found from the result of Prob. 3 above, taken together with the polarizability of a conducting sphere, provided that the distance between charge and sphere is large compared with the radius of the sphere.

5. A dipole of fixed dipole moment is placed in an external electric field. Prove that there is a torque on the dipole proportional to its dipole moment and the magnitude of the electric field, and find how the torque depends on angle.

6. A dipole of fixed dipole moment is placed in an external electric field that is not constant. Prove that there is a force on the dipole depending on the rate of change of field with position, and find how this force depends on the orientation of the dipole and other features of the field.

7. Find the potential as a function of position in a region bounded by surfaces at $x = 0$, $x = L$, $y = 0$, extending to infinity along the y axis, subject to the boundary condition that the potential is zero along the two infinite surfaces $x = 0$, $x = L$, but that it is an arbitrary function of x along the surface from $x = 0$ to $x = L$, $y = 0$. Build up the solution out of individual solutions varying sinusoidally with x, and exponentially with y, noting that they must decrease rather than increase exponentially as y increases.

8. Find the potential as a function of position within a sphere, if the surface of one hemisphere is kept at a potential $+V$, the other hemisphere at a potential $-V$.

9. Set up Laplace's equation in cylindrical coordinates, and solve by separation of variables.

10. A hollow pipe of circular cross section is infinitely long, and is grounded. A disk maintained at potential V practically closes the pipe at a certain point, but is insulated from it. Find the potential as a function of position within the pipe.

11. Show that a solution of Laplace's equation in two dimensions can be written as a linear combination of terms of the form $r^n \begin{matrix}\cos \\ \sin\end{matrix} n\theta$, where n can be any positive or negative integer.

12. Given the two-dimensional potential on a circle of radius R about an origin of the form $V(\theta)$ $(0 \le \theta \le 2\pi)$, where $V(\theta)$ is continuous. Using Green's theorem, show that the potential at any point (r,φ) inside the circle is given by

$$V(r,\varphi) = \frac{R^2 - r^2}{2\pi} \int_0^{2\pi} \frac{V(\theta)\, d\theta}{R^2 + r^2 - 2ar \cos (\varphi - \theta)}.$$

This is known as "Poisson's integral." What form would it take for points exterior to the circle?

13. Find the force and torque on a dipole in the field of a point charge. Find the force exerted by the dipole on the point charge, and verify Newton's third law.

CHAPTER IV

DIELECTRICS

A dielectric, or insulator, is a material containing dipoles whose dipole moment is ordinarily proportional to the applied electric intensity. These dipoles can arise physically in two ways. In the first place, the molecules of any material are composed of positive nuclei, surrounded by rapidly moving negatively charged electrons. These electrons move very freely through the atom or molecule; they are prevented from escaping, however, by intense electric fields. When the molecule is placed in an external electric field, the electrons tend to be displaced by the field, much as the free electrons in a metal are displaced. They cannot travel outside the molecule, however; instead, they pile up on the surface of the molecule, resulting effectively in a surface charge, and the net result is that the molecule becomes polarized, or becomes a dipole. The situation is similar to that of the polarization of a conducting sphere in an external field, discussed in Chap. III, Sec. 5. We can even find a value for the polarizability of a molecule, using the formula $\alpha = 4\pi\epsilon_0 R^3$ of that section, and assuming for R a molecular dimension, which is of the right order of magnitude. If there are N atoms or molecules in volume V, so that the number per unit volume is N/V, and if each one acquires a moment $\alpha\mathbf{E}$, there will then be a dipole moment per unit volume of $(N/V)\alpha\mathbf{E}$. This moment is called the "polarization," and is denoted by $\mathbf{P}$; it is a vector function of position, the polarization being in the same direction as $\mathbf{E}$, in an isotropic dielectric.

The other mechanism by which dipoles can be set up is found in certain materials containing polar molecules; that is, molecules that have dipole moments even in the absence of an external field. Thus a chemical substance like HCl, containing a positive and a negative ion, has dipole molecules, the H end of the molecule being positively, the Cl end negatively charged. In the absence of an external field, the molecules in gaseous or liquid HCl will be oriented in random directions, so that even though each molecule has a moment, the average moment per unit volume will be zero. An impressed field, however, tends to orient the molecules, and it can be shown that there is a net

dipole moment resulting from this, which is proportional to the field. Thus this effect, like the other, gives **P** proportional to **E**. There is a difference, however, which allows the effects to be separated. The induced dipoles are independent of temperature. The orientation of permanent dipoles, however, is opposed by temperature agitation, and by kinetic theory it can be shown that the resulting polarization in a given external field is inversely proportional to the absolute temperature. By investigating the temperature variation of the dipole moment, then, we can experimentally separate the two effects, and can find values both for the polarizability of the individual molecules, and for their permanent dipole moments. For our present purposes, however, we need not distinguish between the two sources of dipoles, and need merely assume the existence of a polarization vector **P**, proportional to **E**, the constant of proportionality of course depending on the nature of the dielectric.

1. The Polarization and the Displacement.—Let us consider a surface S bounding a volume V, within a dielectric. We start with the dielectric unpolarized, and then allow it to polarize. In this process, certain charges will have been carried across the surface S, so that, although originally the volume contains no net charge, there will be such a charge after polarization. Let us find this charge, by computing the amount of charge flowing across an element da of the surface in the process of polarization. We may consider the polarization to consist of many dipoles, each of charge q, with displacement **d** (a vector, pointing along the direction of the dipole moment) between the equal and opposite charges. If **n** is the outer normal to V, the displacement of a charge q along **n** will be $\mathbf{d} \cdot \mathbf{n}$. All those charges contained in the small volume of area da, height $\mathbf{d} \cdot \mathbf{n}$, will be carried over da in the process of polarization. If there are N/V dipoles per unit volume, there will have been a charge of $q(N/V)(\mathbf{d} \cdot \mathbf{n})\, da$ in this small volume, so that the total charge carried out over da will be $(N/V)(q\mathbf{d} \cdot \mathbf{n})\, da$. But $(N/V)q\mathbf{d}$ is simply the polarization **P**, the dipole moment per unit volume. Thus the charge carried out over da is $\mathbf{P} \cdot \mathbf{n}\, da$, and the total charge carried out over the whole surface is the surface integral of this quantity.

If ρ' is the resulting charge density within the volume, set up by the displacement of charge, the charge that has been carried out will be $-\int \rho'\, dv$. Thus we have

$$\int_S \mathbf{P} \cdot \mathbf{n}\, da = -\int_V \rho'\, dv. \tag{1.1}$$

Using the divergence theorem, the left side can be rewritten $\int \operatorname{div} \mathbf{P}\, dv$;

since the relation (1.1) must hold for any volume, the integrands must then be equal, or we have

$$\operatorname{div} \mathbf{P} = -\rho'. \tag{1.2}$$

We may now combine this equation with Eq. (3.3), Chap. II, which we may write $\epsilon_0 \operatorname{div} \mathbf{E} = \rho$, where ρ represents the volume density of charge. We must think carefully how to describe the situation, however. We have two types of charge density: the charge that we deliberately place on our conducting or dielectric bodies, and the charge that automatically appears as a result of polarization. We shall call the first sort the "real charge," the second sort the "polarized charge." Both sorts can produce electric intensity **E**, so that, in Eq. (3.3), Chap. II, the charge density that appears should be the sum of real and polarized charge.

It is customary, however, to use the symbol ρ to denote merely the real charge, so that for instance ρ will be zero within an ordinary dielectric. Thus we must replace our equation by

$$\epsilon_0 \operatorname{div} \mathbf{E} = \rho + \rho',$$

which means the same as Eq. (3.3), Chap. II, but is expressed differently because of the different meaning we now assign to ρ. Using (1.2), we then have

$$\epsilon_0 \operatorname{div} \mathbf{E} = \rho - \operatorname{div} \mathbf{P}, \qquad \operatorname{div} (\epsilon_0\mathbf{E} + \mathbf{P}) = \rho.$$

The combination $\epsilon_0\mathbf{E} + \mathbf{P}$ comes into the theory so often and in such an important way that it is given a special name, the electric displacement, and a special symbol, **D**. Thus we write

$$\operatorname{div} \mathbf{D} = \rho, \qquad \mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P}. \tag{1.3}$$

Equation (1.3) is one of the fundamental equations of electromagnetic theory, and is one of Maxwell's equations.

2. The Dielectric Constant.—Since **P** is proportional to **E**, the displacement **D** is also proportional to **E**. We define the ratio of **D** to **E** as the permittivity, and denote it by ϵ. Thus we have

$$\mathbf{D} = \epsilon\mathbf{E}. \tag{2.1}$$

The ratio of ϵ to ϵ_0, which we may call the permittivity of free space, is the dielectric constant, sometimes called the "specific inductive capacity," and is denoted by κ_e. Thus we may write

$$\kappa_e = \frac{\epsilon}{\epsilon_0}, \qquad \mathbf{D} = \kappa_e\epsilon_0\mathbf{E}. \tag{2.2}$$

The ratio of polarization $\mathbf{P}$ to $\epsilon_0\mathbf{E}$ is called the "susceptibility," and is denoted by χ_e. Thus we have

$$\mathbf{P} = \chi_e\epsilon_0\mathbf{E}, \qquad \kappa_e = 1 + \chi_e. \tag{2.3}$$

Clearly the susceptibility is proportional to the polarizability of the molecules, and the number of molecules per unit volume: in fact, since $\mathbf{P} = (N/V)\alpha\mathbf{E}$, we have

$$\chi_e = \frac{N}{V}\frac{\alpha}{\epsilon_0}. \tag{2.4}$$

We thus have means for finding the dielectric constant of a material, if we know the polarizability of its molecules. There is one word of caution to be expressed, however. It turns out that the polarizability of a molecule is affected by the presence of neighboring molecules. Thus in (2.4) we cannot use the value of α that would be obtained for a molecule in the absence of neighbors, and expect the resulting dielectric constant to be correct for a dense dielectric. We shall take up this correction later.

3. Boundary Conditions at the Surface of a Dielectric.—There are two fundamental equations of electrostatics: (1.3) and Eq. (3.4) of Chap. I, or

$$\operatorname{div} \mathbf{D} = \rho, \qquad \operatorname{curl} \mathbf{E} = 0. \tag{3.1}$$

These equations take on special forms in the case commonly met in electrostatics, in which the dielectrics consist of materials each with a uniform dielectric constant, but with surfaces of discontinuity of dielectric constant between the materials. Within a homogeneous dielectric containing no charges, we have $\operatorname{div} \mathbf{D} = \operatorname{div} \epsilon\mathbf{E} = 0$, so that $\operatorname{div} \mathbf{E} = 0$. From $\operatorname{curl} \mathbf{E} = 0$ we conclude as before that $\mathbf{E} = -\operatorname{grad} \varphi$, and we can then find that $\nabla^2\varphi = 0$, so that Laplace's equation is satisfied for the potential, as in the absence of a dielectric.

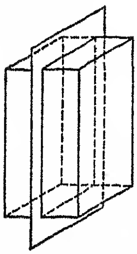


FIG. 9.—Gauss's theorem for normal component of D.

At a surface of discontinuity, however, the situation is quite different. There ϵ is a discontinuous function of position, so that we cannot infer from $\operatorname{div} \epsilon\mathbf{E} = 0$ that $\operatorname{div} \mathbf{E} = 0$; it will not be true. Let us instead investigate an uncharged surface of discontinuity between two dielectrics by Gauss's theorem, using arguments like those in Chap. II, Sec. 1. Over unit area of the surface we erect a thin rectangular volume, as in Fig. 9. We note, from the divergence theorem and the equation $\operatorname{div} \mathbf{D} = \rho$, that Gauss's theorem becomes

$$\int_S \mathbf{D} \cdot \mathbf{n}\, da = \int_V \rho\, dv. \tag{3.2}$$

Since there is no real charge within our volume, this tells us that the total flux of **D** out of the volume is zero; that is, that the flux in over one of the faces equals the flux out over the other, which demands that the normal component of **D** must be the same on both faces. In other words, we conclude that the normal component of **D** is continuous at an uncharged surface of separation between two homogeneous dielectrics.

By an argument similar to that which we have just applied, we may apply Eq. (1.1), rather than (3.2), to the small volume of Fig. 9. Let us suppose that the dielectric is to the left of the surface of separation, empty space to the right. There will then be polarization to the left, and not to the right, so that $\int_S \mathbf{P} \cdot \mathbf{n}\, da$ will be the negative of the component of polarization along the normal pointing away from the dielectric. By (1.1), this equals the polarization charge included within the volume. This must clearly be a surface charge, and since unit area of the surface is included in the volume, we have the result that the normal component of **P**, pointing out of a dielectric, equals the surface polarization charge that appears on the surface as a result of the polarization. It is of course clear that, since the normal component of **D** is continuous at an uncharged surface, and the normal component of **P** is not continuous, the normal component of **E** must likewise be discontinuous.

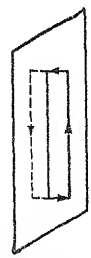


FIG. 10.—Stokes's theorem for surface of discontinuity.

Next we consider the implication of the equation curl $\mathbf{E} = 0$. We draw a contour as shown in Fig. 10, one long side on one side of the surface of separation, the other on the other. By Stokes's theorem in vector analysis, the surface integral of the normal component of the curl of a vector over a surface equals the line integral of the tangential component of the same vector around the contour bounding the surface. Applying this theorem to a surface bounded by the contour in Fig. 10, we note that, since curl **E** is zero, the line integral of the tangential component of **E** about the contour must be zero, which means that the tangential component of **E** must be equal on both long sides of the surface. We then have derived the following general results:

$$\begin{aligned}&\text{From div } \mathbf{D} = \rho\text{: normal component of } \mathbf{D} \text{ continuous}\\ &\text{From curl } \mathbf{E} = 0\text{: tangential component of } \mathbf{E} \text{ continuous,}\end{aligned} \tag{3.3}$$

where these conditions hold at an uncharged surface of separation between dielectrics.

4. Electrostatic Problems Involving Dielectrics, and the Condenser.—Without the principles derived in this chapter, it is almost impossible to see how to solve electrostatic problems involving dielectrics as well as charges. The dielectrics acquire polarized charges on their surfaces, and if we do not know the amount and distribution of these charges, we cannot use our ordinary electrostatic principles to compute the field. The problem is not unlike that which we met in electrostatic problems involving conductors, where we also did not know the distribution of surface charges. We solved that problem by obtaining a potential that was a solution of Laplace's or Poisson's equation in the region outside conductors, but satisfied certain boundary conditions at the surface of a conductor: the potential had to reduce to an appropriate constant on the surface of each conductor, which implied that the field **E** was normal to the surface, or that the tangential component of **E** was zero outside this surface. We see that this condition was a special case of (3.3): since **E** is zero within a conductor, the tangential component of **E** must be zero outside, by continuity. Equation (3.3) gives no information about the normal component at the surface of a conductor, for the surface is not uncharged.

We now notice that a method essentially similar to this one can be used in solving an electrostatic problem involving dielectrics. Inside each dielectric, provided that its dielectric constant is uniform, the potential will satisfy Laplace's or Poisson's equation, where the charge density involved is the real charge, which we assume we know about. At each surface of discontinuity, the conditions (3.3) must be satisfied by the appropriate components of **D** and **E**. Combining with the relation (2.1), $\mathbf{D} = \epsilon\mathbf{E}$, the problem is determined. As a first almost trivial example, consider the field of certain charges embedded in an infinite dielectric of permittivity ϵ. By Gauss's theorem, $\int \mathbf{E} \cdot \mathbf{n}\, da = (1/\epsilon)\int \rho\, dv$. This differs from the corresponding equation for free space only in substituting ϵ for ϵ_0. Thus Green's solution of Poisson's equation will carry through just as in Chap. II, Sec. 4, with the one difference that ϵ will appear in the denominator rather than ϵ_0. Thus the potential, and the field **E**, of a set of charges in a uniform dielectric, will be just $1/\kappa_e$ as large as if the same charges were located in empty space. We notice, however, that the displacement **D** of a set of charges in a uniform dielectric will be the same as if they were in empty space.

As a next example of electrostatic problems involving dielectrics,

consider a condenser filled with a uniform dielectric of permittivity ϵ, rather than with empty space. We can use the same solution of Laplace's equation to represent the potential that we did for the case without dielectrics; the only difference is in the relation between the charge on the plates and the potential difference. For a given voltage between the plates, there will be the same field **E** everywhere in the case with dielectric that there was without it. **D** will, however, be κ_e times as great. By Gauss's theorem, the surface charge at a charged surface will equal D, so that the surface charge on the plates will be κ_e times as great as if the dielectric were absent. The capacity, being the charge divided by the voltage, will then be κ_e times as great as if the condenser had no dielectric but empty space. It is interesting to notice that, at the surface between the dielectric and the conducting plate, there will be not only the real surface charge D, but also be a polarization surface charge, equal to $-P$, or $-(\kappa_e - 1)\epsilon_0 E$. The total charge, real and polarization, on the surface, will then be $\epsilon_0 E$, the same as the total charge for the condenser in the absence of dielectric. The reason a dielectric increases the capacity of a condenser, or the charge that its plates carry for a given voltage, is that a fraction $(\kappa_e - 1)/\kappa_e$ or $1 - 1/\kappa_e$ of the charge is effectively canceled by polarization charge, leaving only the fraction $1/\kappa_e$ for producing a field. Thus, to produce a given field, we must have κ_e times as much charge as in the absence of dielectric.

5. A Charge outside a Semi-infinite Dielectric Slab.—In both the problems we have so far considered, the infinite dielectric and the dielectric filling a condenser, the effect of the dielectric is very simple: the field distribution is just as in empty space, but a given charge produces a field only $1/\kappa_e$ as strong as in empty space. If this situation were always true, the problem of dielectrics would be trivial, and we should not have had to go through all the theory that we have presented in this chapter. We shall now give two examples to show that in general the whole problem is entirely different in the presence of dielectrics from what it is without them. First let us take the problem of a charge q, in empty space, a distance d from an infinite plane surface bounding a semi-infinite dielectric of permittivity ϵ. We can solve this problem by an application of the method of images, similar to that in Chap. I, Sec. 4, but more complicated. In Fig. 11 we shall try to satisfy our conditions by the following assumptions: in the free space, to the left of the surface of separation, we assume that the potential is $q/4\pi\epsilon_0 r_1 - q'/4\pi\epsilon_0 r_2$, where r_1 is the distance from the charge to the point where we are finding the potential, r_2 the distance

from the image to the point, and where the second term is the potential of an image charge $-q'$, whose magnitude we must still determine. In the dielectric to the right of the plane, we assume that the potential is $q''/4\pi\epsilon r_1$, where q'' is also to be determined. These assumed potentials satisfy the first requirement: in the empty space the potential satisfies Poisson's equation, being determined from the charge q; in the dielectric it is determined by Laplace's equation, the charge producing the field being located outside the dielectric. Clearly we cannot have a term varying inversely as r_2, in the dielectric, for there is no real charge located at the image of q.

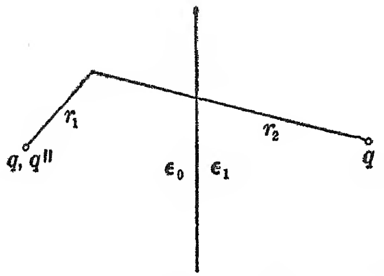


FIG. 11.—Images for point charge outside dielectric slab.

We must now try to satisfy the boundary conditions on the surface. For the tangential component of **E**, we note that the tangential derivative of potential must be the same on both sides of the surface, which means that the potential itself is continuous; this may be used as a substitute for the condition on the continuity of the tangential component of **E**. Thus, remembering that $r_1 = r_2$ at points of the surface, we have

$$q - q' = \frac{q''}{\kappa_e}. \tag{5.1}$$

For the normal component of **D**, we note that the charges q and $-q'$ will produce fields that add rather than subtract. Remembering that we find **D** by taking the gradient of the potential and multiplying by the dielectric constant, this results in

$$q + q' = q''. \tag{5.2}$$

Since (5.1) and (5.2) can be solved for q' and q'' in terms of q, our conditions are compatible, and our assumed potentials give a solution of the problem. Solving for q' and q'', we have

$$q' = \frac{\kappa_e - 1}{\kappa_e + 1}q, \qquad q'' = \frac{2\kappa_e}{\kappa_e + 1}q. \tag{5.3}$$

Lines of force for this problem are shown in Fig. 12.

We can easily check two special cases of (5.3). First, if $\kappa_e = 1$, so that the whole space is really empty, we have $q' = 0$, $q'' = q$, so

that the field is just that of the charge q in empty space. Secondly, if κ_e is infinite, $q' = q$, and $q'' = 2q$. The field $\mathbf{E}$ in the empty space in this case is that produced by the charge q, and an image $-q$; the field within the dielectric is zero, being a field of a finite charge in a medium of infinite dielectric constant. Thus the field in this case is everywhere the same that we should have for a charge q outside a perfect conductor. There are as a matter of fact many ways in which a medium of infinite dielectric constant resembles a conductor.

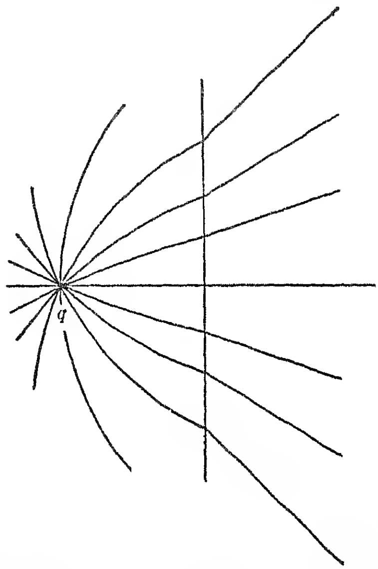


FIG. 12.—Lines of force for point charge outside dielectric slab.

6. Dielectric Sphere in a Uniform Field.—As another example of problems involving dielectrics, we consider a dielectric sphere in a uniform external field, in empty space. Following Eqs. (4.1) and (4.2) of Chap. III, we try the assumption

$$\varphi = \left(-E_0 r + \frac{b}{r^2}\right) \cos \theta$$

outside the sphere, where E_0 is the field at large distances, and where b is to be determined. This assumption correctly gives the field at large distances, and it satisfies Laplace's equation; if b can be chosen to satisfy the boundary conditions at the surface of the sphere, or radius R, we shall have a solution of our problem. Let us also try the assumption that the field is uniform, and equal to E_1, within the sphere. That is, the potential within the sphere is $-E_1 r \cos \theta$. As in the preceding problem, we may take as our two boundary conditions the continuity of the potential, and of the normal component of $\mathbf{D}$. For the potential to be continuous, we must have

$$\left(-E_0 R + \frac{b}{R^2}\right) \cos \theta = -E_1 R \cos \theta, \qquad \text{or} \qquad -E_0 R + \frac{b}{R^2} = -E_1 R. \tag{6.1}$$

For the continuity of the normal component of $\mathbf{D}$, we note that the direction of increasing r is the normal direction. Thus inside the sphere the normal component of $\mathbf{D}$ is $-\epsilon_1 \partial\varphi/\partial r = \epsilon_1 E_1 \cos \theta$, and

outside it is $(\epsilon_0 E_0 + 2\epsilon_0 b/R^3)\cos\theta$. Equating, we have

$$\epsilon_1 E_1 = \epsilon_0 E_0 + \frac{2\epsilon_0 b}{R^3}. \tag{6.2}$$

Solving (6.1) and (6.2) simultaneously, we find

$$E_1 = \frac{3\epsilon_0}{2\epsilon_0 + \epsilon_1} E_0, \qquad 4\pi\epsilon_0 b = 4\pi\epsilon_0 R^3 \frac{(\epsilon_1 - \epsilon_0)}{2\epsilon_0 + \epsilon_1} E_0. \tag{6.3}$$

Thus we have found the complete solution of our problem. The quantity $4\pi\epsilon_0 b$ is the dipole moment of the dipole that would produce the same field at external points as the polarized dielectric sphere. It is interesting again to consider the two limiting cases $\epsilon_1 = \epsilon_0$, and $\epsilon_1 = \infty$. In the first case, we find $E_1 = E_0$, $b = 0$, so that the field is uniform everywhere, as of course it must be if the dielectric sphere is not really present. In the second case, we find $E_1 = 0$,

$$4\pi\epsilon_0 b = 4\pi\epsilon_0 R^3 E_0,$$

so that again the infinite dielectric constant gives the same fields as for a conductor, as discussed in Chap. III, Sec. 4.

There are several remarks to be made about the solution we have found. In the first place, there are only very few problems in which we have the simple situation found here, that the field inside the dielectric object is uniform. This situation holds, in fact, only for an object of ellipsoidal shape. The general ellipsoid can be solved exactly, as well as the sphere, but by considerably more advanced methods than that used here. In the second place, our solution holds equally well for the field within an empty spherical hole in the dielectric of permittivity ϵ_1. As we see from the derivation, to get this result we need only interchange ϵ_0 and ϵ_1 in the solution (6.3). It is interesting to consider the field E_1 within the cavity. The ratio E_1/E_0 can be written

$$\frac{E_1}{E_0} = 1 + \frac{\epsilon_1 - \epsilon_0}{2\epsilon_1 + \epsilon_0} = \frac{\epsilon_1}{\epsilon_0}\left[1 - \frac{2(\epsilon_1 - \epsilon_0)}{2\epsilon_1 + \epsilon_0}\right]. \tag{6.4}$$

Thus the field inside the cavity is greater than in the dielectric [since $(\epsilon_1 - \epsilon_0)/(2\epsilon_1 + \epsilon_0)$ is greater than zero], but is less than ϵ_1/ϵ_0 times the field in the dielectric.

7. Field in Flat and Needle-shaped Cavities.—In the preceding section we have found the field within a spherical cavity in a dielectric. It is interesting also to find the field in a disk-shaped cavity whose normal points along the field, and in a needle-shaped cavity pointing

along the field. In the flat cavity, by symmetry, the field inside as well as outside the cavity will be parallel to the field at a large distance, so that, by the continuity of the normal component of **D**, the value of **D** within the cavity will be equal to that in the dielectric. That is, the value of **E** within the cavity will be ϵ_1/ϵ_0 times the value of **E** in the dielectric, or will be κ_e times the value of **E** in the dielectric. On the other hand, in a needle-shaped cavity, because of the symmetry, the field will again be everywhere parallel, and because of the continuity of the tangential component of **E**, the value of **E** within the cavity will equal that in the dielectric.

These facts are sometimes used to give definitions of **E** and **D** in a dielectric, in a form that could be made the basis of an experimental method of measuring them: **E** is the force on unit charge in a long needle-shaped cavity parallel to the field, and **D** is ϵ_0 times the force on unit charge in a flat disk-shaped cavity with its normal parallel to the field. These definitions are analogous to these introduced in a similar way by Lord Kelvin for defining the corresponding magnetic vectors. It is interesting to note that these two cavities are the limiting forms of ellipsoidal cavities as they get very flat or very elongated, and the solution for the field in an ellipsoidal cavity, which we have mentioned in the preceding section, reduces to these two values in the two limits. In the intermediate case of the sphere, we have seen in the preceding section that the field in the cavity is intermediate between these two limiting values.

Problems

1. A line charge of linear density σ is placed in a medium of dielectric constant $\kappa_1 = \epsilon_1/\epsilon_0$ parallel to and at a distance a from the plane boundary with another medium of dielectric constant $\kappa_2 = \epsilon_2/\epsilon_0$. Find the potential in both media, and show that the force per unit length acting on the line charge is given by

$$\frac{\sigma^2}{4\pi\epsilon_1 a}\frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2}.$$

2. A long circular cylinder of radius a and permittivity ϵ is placed with its axis perpendicular to a uniform electric field E_0 in air. Find the potential inside and outside the cylinder.

3. Find the capacity of a spherical condenser consisting of a conducting sphere of radius r_0, a dielectric of dielectric constant κ_1 for $r_0 < r < r_1$, a dielectric of dielectric constant κ_2 for $r_1 < r < r_2$, and a conducting sphere of radius r_2.

4. Find the capacity per unit length of a cylindrical condenser consisting of a conducting cylinder of radius r_0, a dielectric of dielectric constant κ_1 for $r_0 < r < r_1$, a dielectric of dielectric constant κ_2 for $r_1 < r < r_2$, and a conducting cylinder of radius r_2.

5. Show that, when a line of force cuts through a surface separating two dielectrics of dielectric constants κ_1 and κ_2, it makes angles θ_1 and θ_2 with the normal to the surface in the two media, given by the relation $\kappa_1 \cot \theta_1 = \kappa_2 \cot \theta_2$.

6. A charge q is placed a distance d in front of a semi-infinite dielectric slab of dielectric constant κ_e. Find the force attracting the charge to the slab.

7. A hollow dielectric sphere is placed in a uniform external field. Find the field both inside and outside the sphere, and in the dielectric, in terms of the inner and outer radii of the spherical shell, and its dielectric constant.

CHAPTER V

MAGNETIC FIELDS OF CURRENTS

In Chap. I, we stated that the force on a charge q, moving with a velocity $\mathbf{v}$ in a magnetic field, in which the magnetic induction is $\mathbf{B}$, is $q(\mathbf{v} \times \mathbf{B})$. Similarly we stated that the force on a charge q in an electric field $\mathbf{E}$ was $q\mathbf{E}$. So far we have been studying the electrostatic case. We began by stating the value found experimentally for the electric field $\mathbf{E}$ produced by a charge q. Combination of that statement with the law of force allowed us to get the force between two charges. We shall now begin a similar treatment of magnetic forces. Magnetic fields can be produced, and magnetic forces experienced, by two types of bodies: by charges in motion, or electric currents, and by magnetized bodies, such as permanent magnets. In some cases, the magnetization of magnetic bodies arises from the motion of currents within the atoms, coming from electronic motions, but in other cases it comes from magnetization that is an intrinsic part of the structure of the nuclei or electrons of which the atoms are composed. We could start our study from the fields either of magnets or of currents. We choose to start with currents, as being in a way more fundamental. We shall then show that a current flowing in a small closed path is equivalent to a magnetic dipole, and forms a model for the dipoles present in magnetic materials, which we shall take up in the next chapter.

The magnetic field produced by a steady current, which does not vary with time, is much simpler than that of a current that varies with time, for in the latter case we can have radiation. The problem of magnetic fields of steady currents is called "magnetostatics." It is clear that a single charge in motion cannot produce a static magnetic field, for by its very motion it is found at different points of space at different times. Thus we cannot strictly state the field produced by a charge in motion. Nevertheless, if the charge forms part of a steady current, as if there were a procession of charges, following one after the other in a formation independent of time, then there is a simple law for the field it produces. This law, elucidated by the work of Oersted, of Ampère, and of Biot and Savart, in the

early days of the nineteenth century, may be made the basis of the treatment of magnetostatics, just as the law of Coulomb may be made the basis of electrostatics.

1. The Biot-Savart Law.—It is found experimentally that the magnetic induction resulting from a charge q, moving with a velocity $\mathbf{v}$, at a distance $\mathbf{r}$ away from the charge (where $\mathbf{r}$ is a vector pointing from the charge to the point where the field is being found) is

$$\mathbf{B} = \frac{\mu_0}{4\pi} \frac{q(\mathbf{v} \times \mathbf{r})}{|\mathbf{r}|^3}, \tag{1.1}$$

provided that the moving charge forms part of a current distribution independent of time. In this expression, $\mathbf{B}$ is the magnetic induction in webers per square meter. Since this is not a familiar unit, we can state that 1 weber/sq m equals a magnetic field of 10^4 gausses, in the more usual units. The quantity μ_0 is

$$\mu_0 = 4\pi \times 10^{-7} \text{ henry/m}, \tag{1.2}$$

where we shall find the physical significance of this number later. The charge q is in coulombs, the velocity in meters per second (a vector), and $\mathbf{r}$ is in meters. By virtue of the vector product in (1.1), we see that the lines of $\mathbf{B}$ (that is, lines parallel to the vector $\mathbf{B}$) form circles in a plane normal to the velocity vector $\mathbf{v}$, with centers at the intersection of that vector and the plane, and in such a sense that a right-hand screw advancing along $\mathbf{v}$ would rotate in the direction of $\mathbf{B}$. If θ is the angle between $\mathbf{v}$ and $\mathbf{r}$, the magnitude of $\mathbf{B}$ is equal to

$$|\mathbf{B}| = \frac{\mu_0}{4\pi} \frac{q|\mathbf{v}| \sin \theta}{|\mathbf{r}|^2}.$$

We are often interested not in the field of a moving charge but in that of an element of current, as a length ds of wire carrying a current i. We can easily set up an equivalence between the quantity $q\mathbf{v}$ and a corresponding quantity for the current element. Suppose the cross section of the wire is A, and the charge density of charge in the wire is ρ. Assume this charge is moving with an average velocity v. Then the charge crossing any cross section per second is ρvA. On the other hand, if the current flowing is i, measured in amperes, then i must be the charge crossing a cross section per second. Thus we have $\rho vA = i$. If we multiply both sides by ds, the length of the section of wire, we shall have on the left $\rho A\ ds\ v$, which is the product of the charge density, the volume, and the velocity. That is, it is the total charge contained in the element of wire, times its velocity. This then

equals $i\,ds$. We may make this into a vector equation by letting **v**, and **ds**, both be vectors along the velocity of the charge, or the direction of the wire. Thus we have

$$q\mathbf{v} = i\,\mathbf{ds},$$

so that the magnetic induction resulting from a length **ds** of wire carrying a current i is

$$d\mathrm{B} = \frac{\mu_0}{4\pi}\frac{i(\mathbf{ds} \times \mathbf{r})}{|\mathbf{r}|^3}. \tag{1.3}$$

The law described by (1.1) or (1.3) is often called the "Biot-Savart law." As a result of it, and Eq. (1.1) of Chap. I, for the force acting on a moving charge, we can at once find the force between two elements of current, $i_1\mathbf{ds}_1$ and $i_2\mathbf{ds}_2$, flowing in different conductors. This force is

$$d\mathrm{F} = \frac{\mu_0}{4\pi}i_1 i_2\frac{[\mathbf{ds}_1 \times (\mathbf{ds}_2 \times \mathbf{r})]}{|\mathbf{r}|^3}, \tag{1.4}$$

where we have found the force exerted by the second current element on the first, and where **r** is the vector from the second to the first. This law (1.4) is the equivalent, for current elements, of Coulomb's law, Eq. (2.2) of Chap. I, expressing the law of force between static charges.

The law of force (1.4) serves essentially to define the ampere, the unit of current. Thus we take two current elements, each of unit length, and at unit distance apart, so oriented that the factor

$$\frac{[\mathbf{ds}_1 \times (\mathbf{ds}_2 \times \mathbf{r})]}{|\mathbf{r}|^3}$$

in (1.4) equals unity. Then the currents are defined to be an ampere if the resulting force is 10^{-7} newtons. From this it follows that $\mu_0 = 4\pi \times 10^{-7}$, as in (1.2), a relation that is now seen to be the result of definition rather than of experiment. This on the other hand is not the case with ϵ_0, which we encountered in Eq. (2.1), Chap. I, and which we found from Coulomb's law. The difference is that we define the unit of current, the ampere, from the Biot-Savart law; we define the coulomb, the unit of charge, as the charge flowing per second in a current of 1 amp; and therefore we are not free to choose the constant in Coulomb's law at will, but must determine it by experiment, with the result given in Eq. (2.1), Chap. I. These relations are discussed in more detail in Appendix II.

2. The Magnetic Field of a Linear and a Circular Current.—We can integrate the Biot-Savart law (1.3) to get the magnetic induction B

resulting from any arbitrary steady current. Naturally in practice we must use such an integrated value to determine the ampere, since the interaction between two elements of current, as considered in (1.4) or in the preceding paragraph, cannot be separated from the forces resulting from other current elements in the circuit.

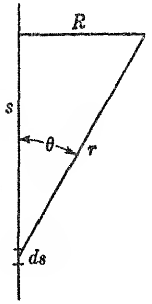


FIG. 13.—Magnetic field from an infinite straight wire.

First we consider the field of an infinite straight wire carrying a current i. From Fig. 13 we see that $r^2 = R^2 + s^2$, where R is the distance from the wire to the point where we are finding the field, and that $\sin\theta = R/\sqrt{R^2 + s^2}$. Thus we have

$$B = \frac{\mu_0}{4\pi} iR \int_{-\infty}^{\infty} \frac{ds}{(R^2 + s^2)^{3/2}} = \frac{\mu_0 i}{2\pi R}. \tag{2.1}$$

The direction of **B** is, of course, in circles surrounding the wire.

Next we consider the field of a circular wire carrying a current, at points along the axis of the wire. The contribution to the field produced by an element ds of the wire, at a point whose distance from the plane is x, will be $(\mu_0/4\pi)i\,ds/(x^2 + R^2)$, as we see in Fig. 14. This field points at an angle to the perpendicular to the plane, however, and only the component along that direction contributes to the resultant field. Thus we must multiply by the factor $R/\sqrt{x^2 + R^2}$ to get the resultant field, which is, after we replace the element ds by the circumference $2\pi R$,

$$B = \frac{\mu_0}{4\pi} i \frac{2\pi R^2}{(x^2 + R^2)^{3/2}} = \frac{\mu_0}{2} i \frac{R^2}{(x^2 + R^2)^{3/2}}. \tag{2.2}$$

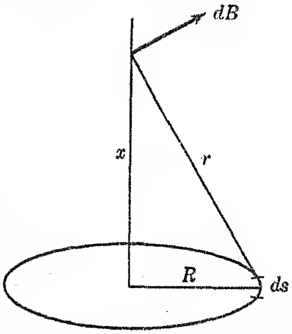


FIG. 14.—Magnetic field from circular current.

We can use the solution (2.2) to find the field in the center of a circular current; it is $\mu_0 i/2R$. This solution is sometimes used to provide an experimental determination of the ampere, following the procedure discussed at the end of the preceding section. It is quite a complicated problem to find the field off the axis in the case of a circular wire.

The solution (2.2) can be used to find the field along the axis of an infinite solenoid, carrying n turns of wire per unit length. To find this we need merely multiply (2.2) by $n\,dx$, and integrate over x, taking account of the turns of the solenoid located at all distances x

from a given point on the axis. Thus we have

$$B = \frac{\mu_0}{2} ni \int_{-\infty}^{\infty} \frac{R^2}{(x^2 + R^2)^{3/2}} dx = \mu_0 ni. \tag{2.3}$$

We shall shortly prove that this field is found everywhere inside the infinite solenoid, not merely along the axis.

3. The Divergence of B, and the Scalar Potential.—In our study of electrostatics, we first found simple methods of obtaining the field by integrating the contributions resulting from all point charges according to Coulomb's law. We soon found, however, that this was a method of restricted usefulness. The really powerful tools of electrostatics come only when we apply general analytical methods, introducing the potential, Gauss's theorem, Poisson's equation, and similar relations. Here too we shall introduce general analytical methods. As a first step, we investigate the divergence and the curl of **B**. We shall first show that div **B** is always zero; then we shall show that curl **B** = 0 in important cases, in which we can therefore introduce a potential, but that this is not true in general, so that we have to adopt other ways of writing **B** in terms of a potential, leading to the concept of a vector potential.

We first consider the divergence of **B**. The lines of **B** form closed circles around the axis of a current element, and since the divergence of a vector signifies the starting or stopping of the vector, we should expect that div **B** = 0. We can prove this, directly from (1.1) or (1.3), and if the theorem holds for the field of an element of current, it must hold for the sum or integral of such fields, or for the field of a complete circuit. We can prove our result by use of the vector relation $\text{div}\,(\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot \text{curl}\,\mathbf{a} - \mathbf{a} \cdot \text{curl}\,\mathbf{b}$, where **a** and **b** are any vectors. We let $\mathbf{a} = \mathbf{v}$, $\mathbf{b} = \mathbf{r}/|\mathbf{r}|^3$, in (1.1). Since **v**, or **a**, is a constant, the first term will be zero. Then we have

$$\text{div}\,\mathbf{B} = \frac{\mu_0}{4\pi} q \left(-\mathbf{v} \cdot \text{curl}\,\frac{\mathbf{r}}{|\mathbf{r}|^3} \right). \tag{3.1}$$

But the vector $\mathbf{r}/|\mathbf{r}|^3$ is simply a vector in the direction of **r**, of magnitude $1/r^2$; that is, it is just like the field of a point charge in electrostatics, whose curl has been proved to be zero in earlier chapters. Thus, from (3.1), we have

$$\text{div}\,\mathbf{B} = 0. \tag{3.2}$$

Equation (3.2) is one of the fundamental equations of electromagnetic

theory, one of Maxwell's equations, which we shall not have to alter in our later development.

Next we consider the curl of **B**. If current flows steadily in a closed loop of wire, we can show that the curl of the resulting magnetic induction is zero everywhere outside the wire, and hence that we can introduce a potential to describe the problem. We shall prove this theorem by setting up a simple and general expression for the potential. As a first step, we consider the solid angle Ω intercepted by the loop of wire, at the point P where we are finding the magnetic induction **B**. If we make a displacement **du** of P, as shown in Fig. 15, the solid angle will change by an amount $d\Omega$. This is the same as the change $d\Omega$ produced by a displacement $-$**du** of the loop. If we make the latter displacement, the change of solid angle will be the sum of all the elementary changes of solid angle intercepted by the small parallelograms bounded by the vectors **du**, **ds**, as shown in the figure. The projection of one of these parallelograms on the normal to the vector r is $-(\mathbf{du} \times \mathbf{ds}) \cdot \mathbf{r}/|\mathbf{r}|$, so that the corresponding increment of solid angle is $-(\mathbf{du} \times \mathbf{ds}) \cdot \mathbf{r}/|\mathbf{r}|^3 = -\mathbf{du} \cdot (\mathbf{ds} \times \mathbf{r})/|\mathbf{r}|^3$.

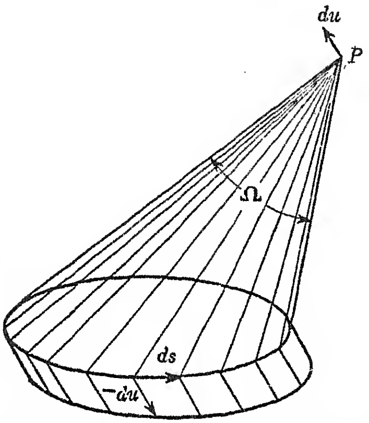


FIG. 15.—Scalar potential of loop of wire carrying current.

Thus the whole change of solid angle, integrating around the current loop, is

$$d\Omega = -\mathbf{du} \cdot \int \frac{\mathbf{ds} \times \mathbf{r}}{|\mathbf{r}|^3}.$$

But at the same time the change $d\Omega$ must be grad $\Omega \cdot$ **du**. Thus we have

$$-\operatorname{grad} \Omega = \int \frac{\mathbf{ds} \times \mathbf{r}}{|\mathbf{r}|^3}.$$

Comparing with (1.3), we see that we have

$$\mathbf{B} = -\frac{\mu_0}{4\pi} i \operatorname{grad} \Omega. \tag{3.3}$$

In other words, the function $(\mu_0/4\pi)i\Omega$ forms a potential, whose

negative gradient gives the magnetic induction **B**. This is generally called a "scalar potential," since it is a scalar quantity, to distinguish it from the vector potential, which we shall soon introduce. It is clear that, in cases in which **B** can be derived from a scalar potential, we must have curl $\mathbf{B} = 0$, since the curl of any gradient is zero.

4. The Magnetic Dipole.—Suppose we have a current i flowing in a positive direction around a very small loop of area A, the normal to the loop being **n**. If θ is the angle between **n** and the vector from the loop to a point P, the solid angle subtended by A at P will be $A \cos \theta / r^2$, so that the scalar potential will be

$$\frac{\mu_0 i A}{4\pi} \frac{\cos \theta}{r^2}.$$

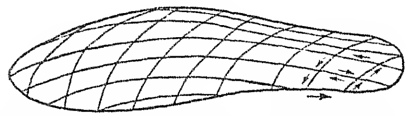


FIG. 16.—Surface made up of small current loops.

This expression depends on position just like the potential (5.1) of Chap. III for an electric dipole. Thus a small loop carrying a current has a magnetic induction **B**, whose lines of force are like those of an electric dipole. We call such a small loop a "magnetic dipole," and define the product iA as the dipole moment. (Some writers define the dipole moment as $\mu_0 i A$ instead, with corresponding change in subsequent formulas involving dipole moments.) If this moment is called m, we have

$$\text{Potential of magnetic dipole} = \frac{\mu_0 m}{4\pi} \frac{\cos \theta}{r^2}. \tag{4.1}$$

In terms of this formulation, there is an interesting way in which we can regard the magnetic field of a current loop as coming from a distribution of magnetic dipoles over a surface spanning the loop. In Fig. 16, we may divide the surface into many small loops, letting a current i flow in each of them, in a positive direction. These currents will cancel each other on all the interior boundaries of the loops, but not over the outer boundary. Thus the sum of the currents of all the small loops will give just the current originally present in the large loop, so that the magnetic field of the small loops must equal that of the large loop. Each of the small loops, of area A, however, produces a field like a dipole of moment iA. In other words, a distribution of dipoles over the surface, or a double layer, whose dipole moment per unit area is i, distributed over any surface spanning the current loop, will have the same magnetic induction at all points that the current loop itself has.

5. Ampère's Law.—Although **B** is derivable from a potential function, we must not assume without further study that the line integral

of **B** about a closed contour is always zero, which we should be inclined to deduce by analogy with Sec. 3, Chap. I. Let us compute the line integral of the tangential component of **B** about a curve as shown in Fig. 17, enclosing the current. Using (3.3), we have

$$\int \mathbf{B} \cdot \mathbf{ds} = -\left(\frac{\mu_0 i}{4\pi}\right) \int d\Omega.$$

If we make a small excursion, and return to our starting point, Ω comes back to its initial value, $\int d\Omega$ is zero, and the line integral $\int \mathbf{B} \cdot \mathbf{ds}$ is zero. If, however, as in Fig. 17, our path encloses the current, the situation is quite different. If we start integrating when we are in the plane of the loop, Ω will be 2π, corresponding to a hemisphere. As we traverse the path in the direction shown, Ω will decrease from this value to zero, become negative, and when we return to our original point it will be -2π. This is on the assumption that solid angles are positive when the point P is above a plane in which the current is circulating in a positive direction, as in Fig. 15. The net change in Ω is then -4π, so that $\int \mathbf{B} \cdot \mathbf{ds} = \mu_0 i$. In other words, when we integrate the tangential component of **B** about a contour enclosing the current i, the integral is $\mu_0 i$; whereas when we integrate about a contour enclosing no current, the integral is zero. The statement of these facts is Ampère's law: the line integral of **B** about a closed contour, in a region containing steady currents, equals μ_0 times the total current flowing through the contour of integration. This general case obviously follows from the derivation we have given, if we regard the whole current flow as being made up of many loops, some of which thread through our contour of integration, some of which do not.

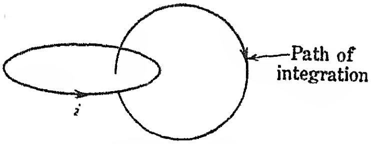


FIG. 17.—Path of integration around current loop.

Ampère's law can often be used, as Gauss's theorem was used in electrostatics, to get the answers to simple problems in the magnetic fields of currents. For instance, we may solve for the field of a linear current, taken up in Sec. 2. If we apply the theorem to a circular contour surrounding the current, **B** is parallel to **ds** by symmetry, and is constant around a circular contour. Thus $B(2\pi R) = \mu_0 i$,

$$B = \left(\frac{\mu_0 i}{2\pi R}\right),$$

as found by direct integration in (2.1). Similarly we can use the theorem to find the field inside a solenoid. Setting up a contour as shown in Fig. 18, we may assume that the field points along the axis by symmetry, and that it is zero outside the solenoid. Then, if the length of the contour is unity, the line integral $\int \mathbf{B} \cdot d\mathbf{s}$ is simply the value of B inside the solenoid. The current enclosed by the contour is ni, if there are n turns in unit length, each carrying current i. Thus we have $B = \mu_0 ni$, as in (2.3). Our present answer is more general, however, for it holds for any point within the solenoid, not merely for points on the axis.

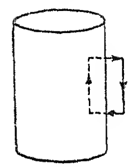


FIG. 18.—Ampère's law for a solenoid.

Ampère's law, which may be stated in the integral form as

$$\int \mathbf{B} \cdot d\mathbf{s} = \mu_0 \Sigma i, \tag{5.1}$$

where Σi indicates the total current threading through the contour, may be written in a differential form by using Stokes's theorem. This theorem of vector analysis states that

$$\int \mathbf{A} \cdot d\mathbf{s} = \int \operatorname{curl} \mathbf{A} \cdot \mathbf{n}\, da,$$

where A is any vector function of position. That is, the line integral of the tangential component of A about a closed contour equals the surface integral of the normal component of curl A over any surface spanning the contour. Using this theorem, we can transform the left side of (5.1). To transform the right side, we introduce the current-density vector J. This represents the number of amperes per square meter flowing in a continuous conductor. The current flowing across a surface element da is then $\mathbf{J} \cdot \mathbf{n}\, da$, so that the total current flowing through the contour is $\int \mathbf{J} \cdot \mathbf{n}\, da$ integrated over a surface spanning the contour. Thus we have as a result of (5.1)

$$\int \operatorname{curl} \mathbf{B} \cdot \mathbf{n}\, da = \mu_0 \int \mathbf{J} \cdot \mathbf{n}\, da.$$

This result must hold for any contour and any surface spanning it; thus the integrands must be equal, and since the integrands are equal for any direction of the vector n, the vectors themselves must be equal, or

$$\operatorname{curl} \mathbf{B} = \mu_0 \mathbf{J}. \tag{5.2}$$

This is the differential formulation of Ampère's law.

We see from (5.2) that, in a region containing no current flow, curl $\mathbf{B} = 0$, but this does not hold within a conductor. In empty space where curl $\mathbf{B} = 0$, we can introduce a potential, as we have done in (3.3), but this cannot be done inside a conductor carrying a current.

The potential we introduced in Sec. 3 for a single loop of current was multiple-valued; that is, as we integrated **B** around a contour enclosing the current the potential increased by $\mu_0 i$, so that if we integrated n times around the contour we should increase the potential by $n\mu_0 i$. In other words, for any single value of the potential, all values obtained by adding or subtracting integral multiples of $\mu_0 i$ are equally legitimate. If now we have many current loops in our space, we can change the potential by integral multiples of μ_0 times any one of the currents, by traversing suitable contours. With enough current loops, this means that we can obtain almost any value of the potential we desire, at a given point of space; and if the current is distributed over the volume, so that we can enclose any amount by traversing a suitable contour within the volume, we can obtain any value of potential whatever at a given point of space. In this case the usefulness of the idea of potential breaks down completely. The integral of **B** is not independent of path, and there is no unique way of defining a scalar potential.

6. The Vector Potential.—It is clear from our discussion of the preceding section that a scalar potential is not useful for discussing a magnetic field, except in the special case where the current flows in a single loop. Mathematically, we have seen in (5.2) that curl **B** is not in general zero, so that the conditions for introducing a potential do not exist. There is, however, another quite different way of setting up a potential function, which is much more general, and equally useful. This is to set up what is called a "vector potential." The fundamental characteristic of a potential is that it is a function that we differentiate to get the vector field we are interested in. The reason why the potential is a useful device in electrostatics is that it is easy to compute it from the charge distribution, by the solution of Poisson's equation given in Eq. (3.5) of Chap. I.

The vector-potential solution for the magnetic field of currents has the same useful features. The existence of a vector potential is based on Eq. (3.2), div $\mathbf{B} = 0$, satisfied by the magnetic induction. By a well-known theorem of vector analysis, div curl $\mathbf{A} = 0$, where **A** is any arbitrary vector; that is, the divergence of any curl is zero. It seems reasonable to assume from this that we can write

$$\mathbf{B} = \text{curl } \mathbf{A}, \tag{6.1}$$

where **A** is a vector function of position, which is the vector potential we have spoken of. By making the assumption (6.1), (3.2) is automatically satisfied, and we must choose **A** so as to satisfy (5.2) as well. This gives us

$$\text{curl curl } \mathbf{A} = \mu_0 \mathbf{J}.$$

By a theorem of vector analysis, curl curl $\mathbf{A}$ = grad div $\mathbf{A} - \nabla^2\mathbf{A}$, where $\mathbf{A}$ is an arbitrary vector. We shall now assume that

$$\operatorname{div} \mathbf{A} = 0. \tag{6.2}$$

We are allowed to make this assumption; it turns out to be the case that, to determine a vector function uniquely, we must specify both its divergence and its curl at all points of space, and (6.1) leaves the divergence undetermined. Making the assumption (6.2), we then have

$$\nabla^2\mathbf{A} = -\mu_0\mathbf{J}. \tag{6.3}$$

That is, $\mathbf{A}$ satisfies Poisson's equation, as φ was shown to in Eq. (3.4), Chap. II. Solving by Eq. (4.3), Chap. II, for an unbounded region, we have the general solution

$$\mathbf{A} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}}{r}\,dv. \tag{6.4}$$

A solution similar to Eq. (6.1) of Chap. III, for a bounded region, can be set up as well. In (6.4) we have a general solution for the vector potential, and hence for the magnetic induction, of any known distribution of currents.

We may show from (6.4) that we are in fact led to the same solution from the Biot-Savart law which we have already discussed, just as the general solution of Poisson's equation led in the electrostatic case to the same result as an elementary discussion of Coulomb's law. In the first place, suppose the current density $\mathbf{J}$ is flowing in a length ds of a conductor of area a. Then $Ja = i$, the current flowing in the conductor, and $J\,dv = Ja\,ds = i\,ds$. Thus the vector potential arising from a current element $i\,\mathbf{ds}$, where we regard $\mathbf{ds}$ as a vector, is

$$d\mathbf{A} = \frac{\mu_0}{4\pi}\frac{i\,\mathbf{ds}}{r}, \tag{6.5}$$

a vector in the direction of the current element. To find the value of $\mathbf{B}$ arising from this vector potential, we take its curl. Noting that $\mathbf{ds}$ is a constant vector, r a scalar function of position, we need to use the formula curl $(f\mathbf{F}) = f$ curl $\mathbf{F}$ + grad $f \times \mathbf{F}$, where f is a scalar, $\mathbf{F}$ a vector function. In our case $\mathbf{F}$ is constant, so that we have

$$d\mathbf{B} = \frac{\mu_0}{4\pi}\,i\operatorname{grad}\left(\frac{1}{r}\right)\times \mathbf{ds}.$$

But grad $(1/r) = -1/r^2$ grad r, and grad $r = \mathbf{r}/|\mathbf{r}|$. Thus finally

$$d\mathbf{B} = \frac{\mu_0}{4\pi}\,i\,\frac{\mathbf{ds}\times\mathbf{r}}{|\mathbf{r}|^3},$$

in agreement with (1.3). We have, then, given a proof of the Biot-Savart law from Eqs. (3.2) and (5.2), and have illustrated our method of handling the vector potential, but have not arrived at a new way of solving the magnetic problem. The solution for a bounded region, similar to Eq. (6.1) of Chap. III, would be new, but it is not of enough practical importance for us to work it out. Although the vector potential has given us no new information in this case, we shall find that in problems involving time variation of charge and current we can still use it, and it then proves a very valuable method of solving for the magnetic induction.

Problems

1. A steady current flows in a circular loop of wire, of radius R. Find the vector potential of the resulting magnetic induction, at large distances compared with R, by adding the contributions to the vector potential due to the separate elements of current.

2. Compute the field, from the potential of the preceding problem, and show that it is approximately that resulting from a single dipole. Find the strength of the dipole, in terms of current and radius R, and check the value derived in the text using the scalar potential.

3. Two parallel straight wires carry equal currents. Work out the magnetic induction due to the two together, in the two cases where the currents flow in the same or in opposite directions, drawing diagrams of the lines of force.

4. Find the magnetic induction at points inside a wire carrying a current, assuming that the wire is straight and of circular cross section and that the current has constant density throughout the wire.

5. Set up the equation for the vector potential in empty space, in cylindrical coordinates, finding components of **A** along r, θ, and z. (*Hint:* Set up the equations curl curl $\mathbf{A} = 0$, div $\mathbf{A} = 0$, using derivatives of the latter, to help simplify the terms.)

6. Use the result of Prob. 5 to find the vector potential of an infinitely long wire carrying a constant current i. From this vector potential find the value of **B**.

7. As in Prob. 5, set up the equation for the vector potential in spherical polar coordinates. Use this result to find the vector potential of a magnetic dipole, comparing with the result of Prob. 1.

8. Two long parallel conductors of circular cross section, each of radius b, are separated by a distance $2a$ between their axes of rotation. If they carry equal and opposite steady currents i, find expressions for the vector potential at all points of a plane (the xy plane) perpendicular to the direction of the currents. From this obtain the magnetic field at interior points of either conductor and set up the integral for the force per unit length with which they repel each other. (Use as an origin the point midway between the wires in the xy plane.) Carry the integration as far as you can.

9. Show that, for any current distribution parallel to a fixed line (the z axis), the magnetic induction may be obtained from a vector potential that has only a z component. From this prove that the lines of magnetic force are given by the equation $A(q_1,q_2) = $ constant, where A is the z component of the vector potential, and q_1, q_2 are orthogonal coordinates in a plane normal to the z axis. Apply this to the results of Prob. 8, and construct a field plot for this case.

CHAPTER VI

MAGNETIC MATERIALS

Just as a dielectric contains electric dipoles that contribute to the field, so there are magnetic media that contain magnetic dipoles. These media are of three sorts: diamagnetic, paramagnetic, and ferromagnetic. A diamagnetic medium contains no permanent dipoles, but only dipoles that are induced by an external field. The atoms of a substance contain electrons that are free in a sense to move about inside the atom, somewhat like the charge in a perfect conductor. We have used this property of the electrons in describing the electric polarization of the atom in an electric field. Similarly in a region where the magnetic induction changes with time, currents are induced in the atom, following the general law of electromagnetic induction which we shall take up in the next chapter. These currents circulate about the direction of the magnetic induction, and thus produce magnetic moments, which prove to be in such a direction that they oppose the field already present. A medium that possesses only this dipole moment opposite to **B** is called a "diamagnetic medium."

Some media contain permanent dipoles. In an unmagnetized body, the dipoles are oriented at random, just as permanent electric dipoles are oriented at random in the absence of an external field, but under the action of an external field the dipoles are oriented, resulting in a dipole moment which, as in the corresponding electric case that we discussed in Chap. IV, is proportional to the external field, and inversely proportional to the absolute temperature. This effect is called "paramagnetism." When it exists, it is usually great enough to mask the diamagnetism which is always present, and which has the opposite sign. The permanent magnetic moments responsible for paramagnetism arise in two ways. In the first place, the theory of atomic structure shows that in many cases there is a permanent circulating current of electrons in the atoms, resulting in a magnetic dipole, as shown in Chap. V. Such circulating currents were first postulated by Ampère, and are often called "Amperian currents," but the explanation of their magnitude and nature was first given by the quantum theory. The diamagnetic atoms are those in which some electrons are circulating in one direction, some in the other, in such a way that

their dipole moments cancel. Secondly, the quantum theory shows that an individual electron possesses a magnetic moment, which is just as characteristic a property as its electric charge. In certain ways this moment can be ascribed to a rotation of the charge about an axis, like the rotation of the earth about its axis, resulting thus in a circulating current; but this simple picture of it cannot be entirely justified. In any case, any atom that has unbalanced moments of the spinning electrons will be paramagnetic; but many atoms that contain even numbers of electrons can have equal numbers of electrons oriented in opposite directions, so that the net magnetic moment is zero, and the atoms are diamagnetic, both orbital and spin magnetic moments canceling.

The final type of medium, the ferromagnetic medium, is really an extreme case of paramagnetism. If the permanent dipoles, generally those resulting from electron spin, are very close together in the medium, there proves to be an effect, explainable only on the quantum theory, and called "exchange," which results in a strong tendency for the spins of adjacent atoms or molecules to line up parallel to each other, even in the absence of an external field. Such a parallel orientation can extend, in an unmagnetized body, over volumes of a considerable scale on an atomic order of magnitude, though a small volume by ordinary standards. Such a volume is called a "domain," and an ordinary unmagnetized ferromagnetic body consists of many domains, each with a strong permanent moment, but oriented in different directions. In the presence of an external magnetic field, the domains change the orientation of their permanent moments, lining them up with the external field, until finally with a very large external field the moment reaches a limit when all moments are parallel. This limit is called the "saturation moment." Reversing the field reverses the moments, but there is an effect similar to friction, hindering this reorientation, so that, by the time the external field is reduced to zero, there can still be a considerable moment. This is the origin of permanent magnetism. If the external field is reversed alternately between one direction and then the other, the moment lags behind the field, resulting in the phenomenon of hysteresis.

These properties of ferromagnetic bodies are very complicated, when one tries to investigate them in detail, in contrast to the diamagnetic and paramagnetic bodies, in which the moment is proportional to the field. The ferromagnetic effect decreases with temperature, the individual domains losing their moments at a critical temperature called the "Curie temperature." Above that temperature the body

becomes paramagnetic, but the moment, instead of being proportional to $1/T$, where T is the absolute temperature, is proportional to $1/(T - \theta)$, where θ is the Curie temperature. The physical explanation of this decrease of ferromagnetism as the temperature increases is that the tendency toward orientation of the magnetic moments which lines them up at low temperatures is opposed by thermal agitation at high temperatures.

1. The Magnetization Vector.—In Sec. 4, Chap. V, we defined the dipole moment of a magnetic dipole: if a current i circulates in a loop of area A, the moment is iA. In a magnetic medium, we shall have dipoles distributed through the volume of the material, and we shall define the magnetization vector **M** as the vector sum of the dipole moments in unit volume. In a diamagnetic or paramagnetic medium, **M** will be proportional to **B**, the constant of proportionality being negative for diamagnetism, positive for paramagnetism; in a ferromagnetic medium, the relationship between **M** and **B** will be much more complicated.

We now observe that the existence of a magnetization vector within a medium implies the existence of currents. Suppose as in Fig. 19 that we have a small rectangular volume, with magnetization along one of the axes. If there is uniform magnetization within the volume, we may replace the magnetic effect of the volume by a current circulating about its faces, as shown. If the area normal to the magnetization is A, the height h, then the total dipole moment is MAh. This would be produced by a current Mh circulating about an area A. In other words, we must assume a surface current at the surface of the magnetized volume, numerically equal to the magnetization. We may indicate the direction as well as the magnitude of this surface current density by the vector equation

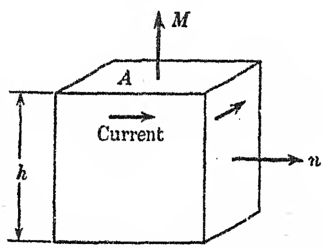


FIG. 19.—Magnetization surface current.

$$\text{Surface current} = \mathbf{M} \times \mathbf{n}, \tag{1.1}$$

where **n** is the outer normal. This is the surface current that appears at a surface of discontinuity between a region with magnetization **M**, and an unmagnetized region.

It is clear that, more generally, there will be a surface current $(\mathbf{M}' - \mathbf{M}'') \times \mathbf{n}$ at a surface of discontinuity between a region

where the magnetization is **M**′, and one where it is **M**″, **n** being the normal pointing from the region **M**′ to that where the magnetization is **M**″. Passing to the limit, there will be a current flowing throughout a volume, when the magnetization varies continuously from point to point within a medium. Let us find the resulting current density. In Fig. 20, we show two adjacent volumes, the upper one being displaced a distance dy along the y axis from the lower one. Let the magnetic moments, in the x direction, be $M'' = M_x(y + dy)$, $M' = M_x(y)$, respectively. Then, considering the contour drawn, the current flowing in it can be written as $J'_z\,dx\,dy$, where $\mathbf{J}'$ is the current density. Thus by the result above we have

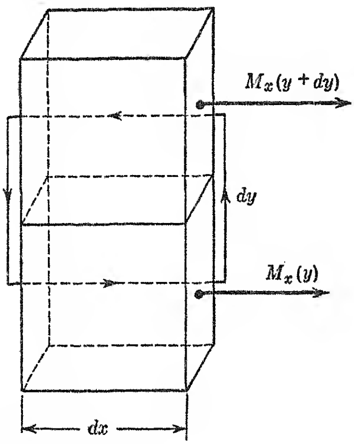


FIG. 20.—Volume magnetization current.

$$[M_x(y + dy) - M_x(y)]\,dx = -J'_z\,dx\,dy.$$

Passing to the limit as dx, dy become small, this becomes

$$J'_z = -\frac{\partial M_x}{\partial y}.$$

We should also have a component of J' in the z direction if M_y varied with x; by a similar argument we should find a contribution $\partial M_y/\partial x$. Thus combining terms we have

$$J'_z = \frac{\partial M_y}{\partial x} - \frac{\partial M_x}{\partial y}.$$

But this is merely the z component of the vector equation

$$\mathbf{J}' = \text{curl}\ \mathbf{M}, \tag{1.2}$$

which is the general relation giving the volume density resulting from a variation of **M** from point to point, a relation of which (1.1) is the limiting form at a surface of discontinuity between a magnetized and an unmagnetized region.

2. The Magnetic Field.—The current density $\mathbf{J}'$ which we have written in (1.2) is a current density that is necessarily present whenever we have a magnetized medium. In our introductory remarks we have seen its interpretation, in terms of the flow of electrons within atoms, and of electron spins. It is quite different from the current

density $\mathbf{J}$ met for instance in Chap. V, Sec. 5, which was a current density resulting from the ordinary flow of current in conductors. The situation is similar to that which we met in Chap. IV, Sec. 1, where we found two types of charge: what we called the "real charge," the charge we placed on the plates of condensers and other conductors, and the "polarized charge," which automatically appeared in dielectrics. There we denoted the polarized charge by a prime, and here similarly we are denoting the current density $\mathbf{J}'$ resulting from magnetic polarization by a prime. We may, if we choose, refer to $\mathbf{J}$ as the real current (that which flows in ordinary conductors), and to $\mathbf{J}'$ as the polarized current, or the magnetization current.

It now follows from our whole train of argument that the magnetization current is just as effective in producing a magnetic induction as a real current. In other words, in Ampère's law, Eq. (5.2) of Chap. V, we must replace $\mathbf{J}$, the current density, by $\mathbf{J} + \mathbf{J}'$, the sum of real and polarized current density. Hence we may rewrite that law

$$\operatorname{curl} \mathbf{B} = \mu_0(\mathbf{J} + \mathbf{J}') = \mu_0(\mathbf{J} + \operatorname{curl} \mathbf{M}),$$

from which we deduce at once the relation

$$\operatorname{curl}\left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M}\right) = \mathbf{J}. \tag{2.1}$$

The quantity $\mathbf{B}/\mu_0 - \mathbf{M}$ which appears in (2.1) is a sufficiently important quantity so that we give it a name and a symbol: we shall call it the "magnetic field," or "magnetic intensity," and denote it by $\mathbf{H}$. The process of introducing it is essentially similar to that used in Chap. IV, Sec. 1, in introducing the electric displacement $\mathbf{D}$. We may rewrite the definition of $\mathbf{H}$ in the form

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}). \tag{2.2}$$

In terms of it, we then have

$$\operatorname{curl} \mathbf{H} = \mathbf{J} \tag{2.3}$$

as the equivalent of (2.1).

In the diamagnetic or paramagnetic medium, we have already stated that $\mathbf{M}$ is proportional to $\mathbf{B}$. Thus both these vectors are also proportional to $\mathbf{H}$, and we may write

$$\mathbf{B} = \mu\mathbf{H} = \kappa_m\mu_0\mathbf{H}, \tag{2.4}$$

where the dimensionless quantity κ_m is generally referred to as the "permeability." The ratio of magnetization to magnetic intensity

is generally called the "magnetic susceptibility":

$$\mathbf{M} = \chi_m \mathbf{H}, \qquad \kappa_m = 1 + \chi_m. \tag{2.5}$$

These equations are closely analogous to Eqs. (2.3) of Chap. IV, though μ_0 does not appear in the same way in which ϵ_0 did in the earlier case. Unlike the case of dielectrics, χ_m can be negative (for diamagnetic media) as well as positive (for paramagnetic media). The magnitudes are also quite different: for diamagnetic and paramagnetic media, κ_m differs by very small amounts from unity, and χ_m is very small compared with unity, whereas the dielectric constant κ_e can be quite large compared with unity. For ferromagnetic media, where **M** is not proportional to **B** or **H**, we can define a permeability by (2.4), but it will be a function of **H**, rather than a constant. It is more convenient for such media simply to give experimental curves for **B** as a function of **H**, as is done in the conventional hysteresis curves, which we shall mention in a later section.

The magnetization vector measures the density of magnetic dipoles per unit volume, just as the polarization **P** measures the density of electric dipoles. In Chap. IV, we found that **P** was associated with a volume and surface-charge density: by Eq. (1.2) of that chapter we found that the volume density ρ' was given by div $\mathbf{P} = -\rho'$, and in Sec. 3 of that chapter we found that at a surface of a dielectric a surface-charge density appeared, equal numerically to the component of **P** along the outer normal. In an analogous way we may define a volume density of magnetic poles, equal to $-$ div **M**, and a surface density equal to the component of **M** along the outer normal to a magnetized body. Remembering that div $\mathbf{B} = 0$, we see that div $\mathbf{H} = -$ div **M**. Thus lines of **H** will diverge outward from regions where the density of poles is positive, or from north poles, and will converge to regions of negative density, or south poles. Lines of **B**, on the other hand, are continuous, on account of the equation div $\mathbf{B} = 0$. We shall shortly see an example of this behavior of magnetic lines, in the case of the magnetized sphere. It should be understood that, although we have an analogy between electric charge and magnetic pole strength, this analogy is far from complete, since electric charges can be separated from each other, whereas magnetic poles can exist only in dipoles, or in polarized bodies.

3. Magnetostatic Problems Involving Magnetic Media.—We have found two fundamental equations dealing with the magnetic field,

$$\text{div } \mathbf{B} = 0, \qquad \text{curl } \mathbf{H} = \mathbf{J}, \tag{3.1}$$

where the first is Eq. (3.2) of Chap. V, and the second (2.3). These,

taken together with (2.4), or the corresponding relation for ferromagnetic media given by the hysteresis curve, give enough information so that we can solve magnetostatic problems involving magnetic media as well as currents. Problems in diamagnetism and paramagnetism can be solved as easily as those involving dielectrics, and by essentially the same methods. We generally have a number of media, each of uniform permeability, with surfaces of separation between. Generally also we do not have any currents $\mathbf{J}$ flowing in the magnetic media. Thus inside these media we have $\operatorname{div} \mathbf{B} = 0$, from which, since $\mathbf{B} = \mu\mathbf{H}$, and μ is constant, we have $\operatorname{div} \mathbf{H} = 0$; furthermore $\operatorname{curl} \mathbf{H} = 0$; $\mathbf{H}$ can then be derived from a scalar potential, which satisfies Laplace's equation because of $\operatorname{div} \mathbf{H} = 0$. Thus the fields inside the various media can be found by familiar methods, and we have only to satisfy boundary conditions at the surfaces of discontinuity, and to ensure that our fields behave properly at the currents $\mathbf{J}$. The boundary conditions, by methods already used, are

From $\operatorname{div} \mathbf{B} = 0$: normal component of $\mathbf{B}$ continuous

From $\operatorname{curl} \mathbf{H} = \mathbf{J}$: tangential component of $\mathbf{H}$ continuous at a surface carrying no current; if there is a surface current, surface-current density $=$ (discontinuity in $\mathbf{H}$) $\times$ $\mathbf{n}$. (3.2)

As a simple example, let us consider the problem of a solenoid filled with a diamagnetic or paramagnetic material. As in Chap. V, Sec. 5, we have a solenoid with n turns per unit length, each carrying a current i. The solenoid is filled with a material of permeability κ_m. The surface-current density is ni; thus the discontinuity in tangential component of $\mathbf{H}$ between the inside and outside of the solenoid is ni. Since $\mathbf{H}$ is zero outside, we see that inside $\mathbf{H} = ni$, and $\mathbf{B} = \mu ni$. In other words, the value of $\mathbf{H}$ within an infinite solenoid is the same, independent of the medium within it, so long as there are the same number of ampere-turns in the winding, but the value of $\mathbf{B}$ is proportional to the permeability of the medium. This problem is one that shows us the proper units to use for measuring $\mathbf{H}$: since $\mathbf{H} = ni$, we measure magnetic intensity in ampere-turns per meter. A problem similar to the solenoid, and conveniently realizable experimentally, is the toroid, a ring-shaped piece of magnetic material, wound round and round with windings, so that the magnetic lines circulate within the ring. Simple application of Ampère's law shows that, just as in the solenoid, the value of $\mathbf{H}$ is equal to the number of ampere-turns per meter in the winding, provided that the cross section

is generally called the "magnetic susceptibility":

$$\mathbf{M} = \chi_m \mathbf{H}, \qquad \kappa_m = 1 + \chi_m. \tag{2.5}$$

These equations are closely analogous to Eqs. (2.3) of Chap. IV, though μ_0 does not appear in the same way in which ϵ_0 did in the earlier case. Unlike the case of dielectrics, χ_m can be negative (for diamagnetic media) as well as positive (for paramagnetic media). The magnitudes are also quite different: for diamagnetic and paramagnetic media, κ_m differs by very small amounts from unity, and χ_m is very small compared with unity, whereas the dielectric constant κ_e can be quite large compared with unity. For ferromagnetic media, where **M** is not proportional to **B** or **H**, we can define a permeability by (2.4), but it will be a function of **H**, rather than a constant. It is more convenient for such media simply to give experimental curves for **B** as a function of **H**, as is done in the conventional hysteresis curves, which we shall mention in a later section.

The magnetization vector measures the density of magnetic dipoles per unit volume, just as the polarization **P** measures the density of electric dipoles. In Chap. IV, we found that **P** was associated with a volume and surface-charge density: by Eq. (1.2) of that chapter we found that the volume density ρ' was given by div $\mathbf{P} = -\rho'$, and in Sec. 3 of that chapter we found that at a surface of a dielectric a surface-charge density appeared, equal numerically to the component of **P** along the outer normal. In an analogous way we may define a volume density of magnetic poles, equal to $-$ div **M**, and a surface density equal to the component of **M** along the outer normal to a magnetized body. Remembering that div $\mathbf{B} = 0$, we see that div $\mathbf{H} = -$ div **M**. Thus lines of **H** will diverge outward from regions where the density of poles is positive, or from north poles, and will converge to regions of negative density, or south poles. Lines of **B**, on the other hand, are continuous, on account of the equation div $\mathbf{B} = 0$. We shall shortly see an example of this behavior of magnetic lines, in the case of the magnetized sphere. It should be understood that, although we have an analogy between electric charge and magnetic pole strength, this analogy is far from complete, since electric charges can be separated from each other, whereas magnetic poles can exist only in dipoles, or in polarized bodies.

3. Magnetostatic Problems Involving Magnetic Media.—We have found two fundamental equations dealing with the magnetic field,

$$\operatorname{div} \mathbf{B} = 0, \qquad \operatorname{curl} \mathbf{H} = \mathbf{J}, \tag{3.1}$$

where the first is Eq. (3.2) of Chap. V, and the second (2.3). These,

taken together with (2.4), or the corresponding relation for ferromagnetic media given by the hysteresis curve, give enough information so that we can solve magnetostatic problems involving magnetic media as well as currents. Problems in diamagnetism and paramagnetism can be solved as easily as those involving dielectrics, and by essentially the same methods. We generally have a number of media, each of uniform permeability, with surfaces of separation between. Generally also we do not have any currents $\mathbf{J}$ flowing in the magnetic media. Thus inside these media we have div $\mathbf{B} = 0$, from which, since $\mathbf{B} = \mu\mathbf{H}$, and μ is constant, we have div $\mathbf{H} = 0$; furthermore curl $\mathbf{H} = 0$; $\mathbf{H}$ can then be derived from a scalar potential, which satisfies Laplace's equation because of div $\mathbf{H} = 0$. Thus the fields inside the various media can be found by familiar methods, and we have only to satisfy boundary conditions at the surfaces of discontinuity, and to ensure that our fields behave properly at the currents $\mathbf{J}$. The boundary conditions, by methods already used, are

From div $\mathbf{B} = 0$: normal component of $\mathbf{B}$ continuous

From curl $\mathbf{H} = \mathbf{J}$: tangential component of $\mathbf{H}$ continuous at a surface carrying no current; if there is a surface current, surface-current density = (discontinuity in $\mathbf{H}$) $\times$ $\mathbf{n}$. (3.2)

As a simple example, let us consider the problem of a solenoid filled with a diamagnetic or paramagnetic material. As in Chap. V, Sec. 5, we have a solenoid with n turns per unit length, each carrying a current i. The solenoid is filled with a material of permeability κ_m. The surface-current density is ni; thus the discontinuity in tangential component of $\mathbf{H}$ between the inside and outside of the solenoid is ni. Since $\mathbf{H}$ is zero outside, we see that inside $\mathbf{H} = ni$, and $\mathbf{B} = \mu ni$. In other words, the value of $\mathbf{H}$ within an infinite solenoid is the same, independent of the medium within it, so long as there are the same number of ampere-turns in the winding, but the value of $\mathbf{B}$ is proportional to the permeability of the medium. This problem is one that shows us the proper units to use for measuring $\mathbf{H}$: since $\mathbf{H} = ni$, we measure magnetic intensity in ampere-turns per meter. A problem similar to the solenoid, and conveniently realizable experimentally, is the toroid, a ring-shaped piece of magnetic material, wound round and round with windings, so that the magnetic lines circulate within the ring. Simple application of Ampère's law shows that, just as in the solenoid, the value of $\mathbf{H}$ is equal to the number of ampere-turns per meter in the winding, provided that the cross section

is small compared with the length of the toroid, and the value of **B** is μ times as great.

Another simple problem is the field produced by any distribution of currents in an infinite space filled with a medium of permeability κ_m. Because of the equation curl $\mathbf{H} = \mathbf{J}$, the magnetic intensity **H** produced by the currents will be independent of κ_m. Thus **B** will be κ_m times as great as in the corresponding problem in empty space. Since the force on another current element is proportional to **B**, we see that the force between two current elements immersed in a medium of permeability κ_m is κ_m times as great as if the same current elements were in empty space. We may take account of this, in Eq. (1.4), Chap. V, for the force between two current elements, by replacing the quantity μ_0 which appears in that expression by μ. We must remember, however, just as in the corresponding case of dielectrics, that the general case of a number of different magnetic media with surfaces of separation between them is a complicated problem, and that it is by no means true that the value of **B** at any point is merely κ_m times as great as it would be if no magnetic media were present.

4. Uniformly Magnetized Sphere in an External Field.—Just as in the corresponding dielectric problem taken up in Sec. 6, Chap. IV, the problem of a uniformly magnetized sphere in an external field is one that can be solved exactly, and that is of considerable physical importance. Let us first consider a sphere of radius R with a uniform magnetization **M** along the axis of spherical coordinates, without an external field (such as we might have from a permanent magnet). We can later superpose a constant external field. We can easily find that the field of this sphere is as follows:

$$\text{Inside the sphere, } \mathbf{H} = -\frac{\mathbf{M}}{3}, \qquad \mathbf{B} = \frac{2}{3}\mu_0\mathbf{M}, \text{ along axis}$$

$$\text{Outside the sphere, } \mathbf{H} = -\operatorname{grad}\left(\frac{MR^3}{3r^2}\cos\theta\right), \qquad \mathbf{B} = \mu_0\mathbf{H}. \quad (4.1)$$

We can verify this solution as follows: As in Sec. 6, Chap. IV, **B** and **H** outside the sphere are the field of a dipole, of moment $\frac{4}{3}\pi R^3 M$ (where we use Chap. V, Sec. 4, to get the moment from the scalar potential). This is simply the moment of a sphere of radius R, with a constant density of magnetization **M**. The field outside the sphere is then a gradient of a scalar potential, so that it is a solution of the equations governing **B** and **H** in empty space. Inside the sphere, the constant values of **B** and **H** are likewise solutions of these equations. We can find the normal components of **B**, and tangential components

of **H**, inside and outside the sphere, at radius R, and show them to be continuous. Finally, inside the sphere we have the relation

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}),$$

satisfying Eq. (2.2).

The lines of **B** inside and outside the sphere are then as given in Fig. 21. Because of the relation div $\mathbf{B} = 0$, the lines are closed, the normal flux being continuous at the surface. On the other hand,

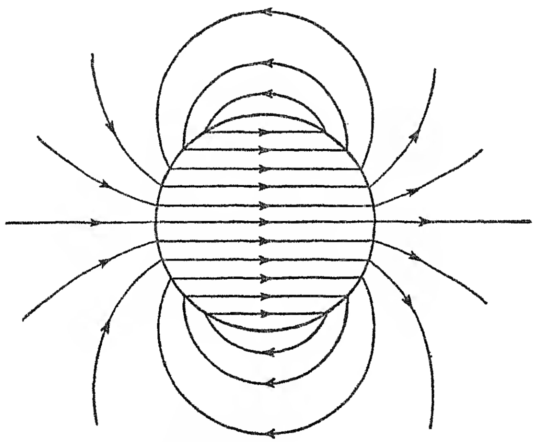


FIG. 21.—Lines of force of uniformly magnetized sphere.

although the lines of **H** outside the sphere are like those of **B**, the lines of **H** inside run in the opposite direction. Thus lines of **H** originate from the right face of the sphere (the north pole of the magnet), and terminate at the left face (the south pole). If now the magnet is in a constant external field $\mathbf{H}_0$, whose corresponding value of **B** is $\mathbf{B}_0$, pointing along the axis, and equal to $\mu_0\mathbf{H}_0$, we need only add these constant values $\mathbf{H}_0$ and $\mathbf{B}_0$ to the solution of (4.1), both inside and outside the sphere, to get the complete solutions. We thus find that inside a magnetized sphere placed in an external field $\mathbf{H}_0$, the magnetic field is only $\mathbf{H}_0 - \mathbf{M}/3$. This effect, by which there is a term $-\mathbf{M}/3$ subtracted from $\mathbf{H}_0$ because of the magnetization, is called the "demagnetizing effect," and the factor $\frac{1}{3}$, which depends on the geometry of the sphere, is called the "demagnetizing factor."

We can now use these relations, together with a hysteresis curve, to investigate the magnetic moment that a ferromagnetic sphere would acquire in a given external field. If B, H, represent the values within the sphere, in an external field H_0, we then have

$$H = H_0 - \frac{M}{3}, \qquad \frac{B}{\mu_0} = H_0 + \frac{2M}{3}, \qquad \frac{B}{\mu_0} = -2H + 3H_0. \tag{4.2}$$

If we know the relation between B and H characteristic of the material (that is, the hysteresis curve), we can then find the intersection of this curve with the curve $B/\mu_0 = -2H + 3H_0$, from (4.2). This intersection will give us the values of B and H inside the sphere in any external field. From those, we can get M, and hence the magnetic moment of the sphere. An example of a hysteresis curve, with the straight lines (4.2) superposed, is given in Fig. 22. Conversely, by a measurement of the magnetic moment, we can work backward and find the hysteresis curve.

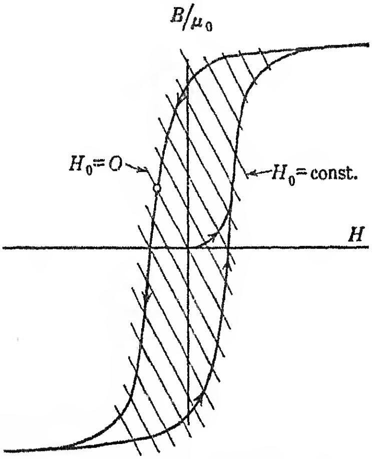


FIG. 22.—Hysteresis curve.

As an example of this method, we may ask what are the values of B and H within the sphere, if we first magnetize to saturation, and then remove the external field. If $H_0 = 0$, we have $B/\mu_0 = -2H$, and the values of B and H are determined by the intersection shown in the figure. We note that this has a much smaller value of B (and hence a much smaller magnetic moment) than if the demagnetizing factor were smaller. For suppose the factor is L, so that

$$H = H_0 - LM,$$

$B/\mu_0 = H_0 + (1 - L)M$. Then we have

$$\frac{B}{\mu_0} = H\left(1 - \frac{1}{L}\right) + \frac{H_0}{L}. \tag{4.3}$$

Thus, as L approaches zero, the slope of the straight line corresponding to $H_0 = 0$ becomes negatively infinite, or the line coincides with the axis of ordinates in the BH diagram, so that the B inside the permanently magnetized object becomes much greater.

This situation can be approached in a practical case. It turns out that we can obtain an exact solution of the problem of a uniformly magnetized body, similar to (4.1), not only for the sphere, but for the ellipsoid; though the solutions are much more complicated, involving elliptic integrals. When the ellipsoid is infinitely elongated along the

axis, the magnet approaches an infinitely thin bar magnet; in this case the demagnetizing factor approaches zero. The opposite limit is the flattened disk-shaped ellipsoid, in which the demagnetizing factor approaches unity, so that B/μ_0 approaches H_0, and in the absence of an external field there is no B within the magnet. Because of the possibility of getting exact solutions, accurate experiments on the magnetic behavior of magnetic materials are generally made on samples of an ellipsoidal shape. As a final example of similar arguments, we can find the values of B and H within ellipsoidal cavities in a magnetized medium, as we did in Sec. 7, Chap. IV, for the corresponding dielectric case. We find that with a long thin cavity the value of H equals the value H_0 in the medium at a large distance from the cavity, whereas with a disk-shaped cavity the value of H equals the value of B_0 in the medium at a large distance, divided by μ_0.

5. Magnetomotive Force.—Practical magnets are not made in the spherical or ellipsoidal shapes discussed in the preceding section, but rather in the shape of a closed ring of some shape with an air gap, and the external field is supplied by windings surrounding the ring, as shown in Fig. 23. In this case, if the permeability of the magnetic medium is high, the lines of force will mostly flow through the magnetic material except near the gap, where the lines of force will partly cross the gap where we are trying to produce a high field, and will partly leak around the sides of that gap. It is almost impossible to calculate accurately the form of the resulting field; we do not have the simplifying feature that the field is constant within the magnetic medium, as we do with the ellipsoidal or spherical magnet. We can, however, make certain assumptions that allow us to use a certain amount of theory in discussing such a magnet.

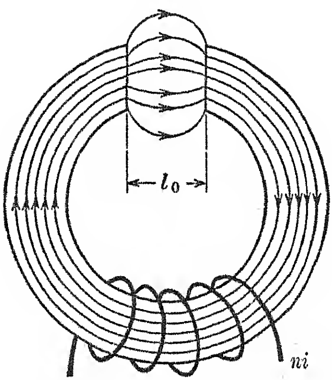


FIG. 23.—Lines of force of ring magnet.

Let us assume that the leakage is concentrated in a relatively short part of the magnet, near the poles. Then, if the cross section is uniform through most of the length, we note that the flux of **B** must be the same through most cross sections (because of the lack of leakage) and hence **B** must be constant along the length of the magnet (because of the uniform cross section). Let this constant value of **B**

$$H = H_0 - \frac{M}{3}, \qquad \frac{B}{\mu_0} = H_0 + \frac{2M}{3}, \qquad \frac{B}{\mu_0} = -2H + 3H_0. \quad (4.2)$$

If we know the relation between B and H characteristic of the material (that is, the hysteresis curve), we can then find the intersection of this curve with the curve $B/\mu_0 = -2H + 3H_0$, from (4.2). This intersection will give us the values of B and H inside the sphere in any external field. From those, we can get M, and hence the magnetic moment of the sphere. An example of a hysteresis curve, with the straight lines (4.2) superposed, is given in Fig. 22. Conversely, by a measurement of the magnetic moment, we can work backward and find the hysteresis curve.

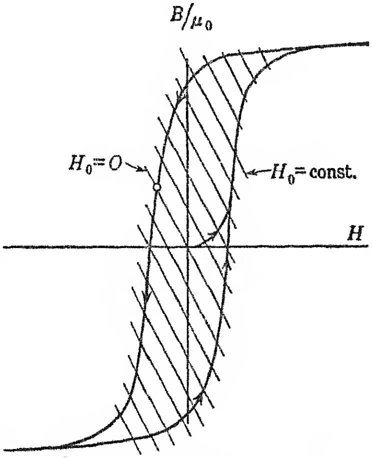


FIG. 22.—Hysteresis curve.

As an example of this method, we may ask what are the values of B and H within the sphere, if we first magnetize to saturation, and then remove the external field. If $H_0 = 0$, we have $B/\mu_0 = -2H$, and the values of B and H are determined by the intersection shown in the figure. We note that this has a much smaller value of B (and hence a much smaller magnetic moment) than if the demagnetizing factor were smaller. For suppose the factor is L, so that

$$H = H_0 - LM,$$

$B/\mu_0 = H_0 + (1 - L)M$. Then we have

$$\frac{B}{\mu_0} = H\left(1 - \frac{1}{L}\right) + \frac{H_0}{L}. \qquad (4.3)$$

Thus, as L approaches zero, the slope of the straight line corresponding to $H_0 = 0$ becomes negatively infinite, or the line coincides with the axis of ordinates in the BH diagram, so that the B inside the permanently magnetized object becomes much greater.

This situation can be approached in a practical case. It turns out that we can obtain an exact solution of the problem of a uniformly magnetized body, similar to (4.1), not only for the sphere, but for the ellipsoid; though the solutions are much more complicated, involving elliptic integrals. When the ellipsoid is infinitely elongated along the

axis, the magnet approaches an infinitely thin bar magnet; in this case the demagnetizing factor approaches zero. The opposite limit is the flattened disk-shaped ellipsoid, in which the demagnetizing factor approaches unity, so that B/μ_0 approaches H_0, and in the absence of an external field there is no B within the magnet. Because of the possibility of getting exact solutions, accurate experiments on the magnetic behavior of magnetic materials are generally made on samples of an ellipsoidal shape. As a final example of similar arguments, we can find the values of B and H within ellipsoidal cavities in a magnetized medium, as we did in Sec. 7, Chap. IV, for the corresponding dielectric case. We find that with a long thin cavity the value of H equals the value H_0 in the medium at a large distance from the cavity, whereas with a disk-shaped cavity the value of H equals the value of B_0 in the medium at a large distance, divided by μ_0.

5. Magnetomotive Force.—Practical magnets are not made in the spherical or ellipsoidal shapes discussed in the preceding section, but rather in the shape of a closed ring of some shape with an air gap, and the external field is supplied by windings surrounding the ring, as shown in Fig. 23. In this case, if the permeability of the magnetic medium is high, the lines of force will mostly flow through the magnetic material except near the gap, where the lines of force will partly cross the gap where we are trying to produce a high field, and will partly leak around the sides of that gap. It is almost impossible to calculate accurately the form of the resulting field; we do not have the simplifying feature that the field is constant within the magnetic medium, as we do with the ellipsoidal or spherical magnet. We can, however, make certain assumptions that allow us to use a certain amount of theory in discussing such a magnet.

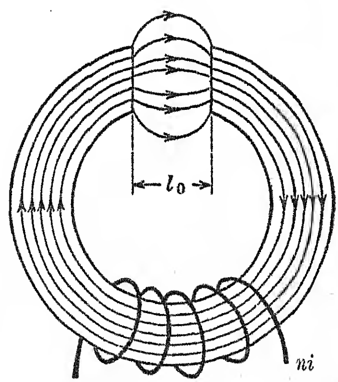


FIG. 23.—Lines of force of ring magnet.

Let us assume that the leakage is concentrated in a relatively short part of the magnet, near the poles. Then, if the cross section is uniform through most of the length, we note that the flux of **B** must be the same through most cross sections (because of the lack of leakage) and hence **B** must be constant along the length of the magnet (because of the uniform cross section). Let this constant value of **B**

within the magnet be B_1, and the corresponding value of $\mathbf{H}$ be H_1. Furthermore let the length of the magnetic material, along a mean circumference, be l_1. In the gap, if there were no leakage, the value of $\mathbf{B}$ would be the same as inside the magnet; because of leakage, however, it will be smaller. Suppose the value of $\mathbf{B}$ in the gap is then $B_0 = \alpha B_1$, where α is a factor less than unity, which is sometimes called the "leakage factor." Then the value of $\mathbf{H}$ in the gap will be

$$H_0 = \frac{B_0}{\mu_0} = \frac{\alpha B_1}{\mu_0}.$$

Let the length of the gap be l_0. Now we apply Ampère's law to a contour through the middle of the magnet. From (2.3), Ampère's law may be stated in the form

$$\int \mathbf{H} \cdot \mathbf{ds} = ni, \tag{5.1}$$

where there are n turns of wire, each carrying i amp, threading the contour. We then have approximately

$$H_0 l_0 + H_1 l_1 = ni. \tag{5.2}$$

We may then rewrite (5.2) in the form

$$\frac{B_1}{\mu_0} = -\frac{l_1}{\alpha l_0} H_1 + \frac{ni}{\alpha l_0}. \tag{5.3}$$

This equation, relating B_1 and H_1, is essentially equivalent to (4.3). In place of the factor $(1/L - 1)$ multiplying H we have $l_1/\alpha l_0$; in place of H_0 we have the term in ni. It is customary to call the number ni, which is the number of ampere-turns, the "magnetomotive force" (mmf), by analogy with the electromotive force $\int \mathbf{E} \cdot \mathbf{ds}$ met in the electric field. As in the preceding section, in a plot of B_1 versus H_1 a line of constant mmf will be a straight line of negative slope. In an actual case, l_1 will ordinarily be large compared with l_0, and α will be rather small. Thus the case resembles that of a small demagnetizing factor, which we have discussed in connection with (4.3), so that the value of B_1 inside the magnet will be large if the mmf is removed, as in a permanent magnet.

We may note a number of interesting consequences of our theory. If the magnet is operated as an electromagnet, with large mmf, the value of B_1 will be inversely proportional to l_0; that is, for a large gap, we shall have to have a correspondingly high mmf to get saturation of the ferromagnetic material, which we presumably want to do to operate the magnet efficiently. A small value of α, on the contrary,

allows us to saturate the material with a smaller mmf, but it gives correspondingly smaller fields in the gap. If the magnet is operated as a permanent magnet, the field in the gap, which we are primarily interested in, is given by

$$H_0 = -\frac{l_1}{l_0} H_1,$$

so that, if the material is used with a given value of B_1 and of H_1, the field in the gap will be inversely proportional to the length of the gap, for constant magnet length l_1. It can be shown, from arguments of energy that we are not yet prepared to understand, that a permanent magnet can be made with least material if it is made with those particular values of B_1 and H_1 which make the product B_1H_1 a maximum; if this condition is applied, the dimensions of a magnet designed for a given field and given gap are determined.

Problems

1. An electron of charge e, mass m, rotates in a circle of radius r with velocity v. Using classical methods, show that the magnetic moment is proportional to the angular momentum. If the angular momentum is $h/2\pi$, where h is Planck's constant, find the magnetic moment, in mks units. If we make a simple model of a ferromagnetic substance by putting one such moment at each point of a simple cubic lattice of lattice spacing 3 A, find the magnetization per unit volume, M, when all moments are parallel.

2. A sphere of radius a carrying a uniform surface-charge density σ is rotated about a diameter with constant angular velocity ω. Calculate the magnitude and direction of the magnetic intensity **H** at the center of the sphere. Show that this value of **H** is the same at all points inside the sphere.

3. A long circular cylinder of permeability κ_m is placed in a uniform external magnetic field that is perpendicular to the generators of the cylinder, resulting in a uniform magnetization of the cylinder. Find expressions for **B** and **H** inside and outside the cylinder, and the demagnetizing factor.

4. A long straight wire carrying a steady current i is placed parallel to and at a distance a from a large thick plane slab of permeability κ_m. Show that the field inside the slab is that which would be produced by a current $2i/(\kappa_m + 1)$ in the same wire embedded in an infinite medium of permeability κ_m. Show further that the field in the space in front of the slab is that produced by a current i and another current $i(\kappa_m - 1)/(\kappa_m + 1)$ parallel to and at a distance a behind the interface, both in empty space. What is the force per unit length acting on the current-carrying wire?

5. Discuss the magnetic field of permanent magnets, showing that **H** may be obtained as the negative gradient of a single-valued scalar potential ψ, which satisfies Poisson's equation $\nabla^2\psi = -\rho'$, with $\rho' = \operatorname{div} \mathbf{H}$. If one defines magnetic-pole strength as $p = \int \rho' \, dv$, show that two concentrated poles exert mechanical forces on each other given by $F = \frac{p_1p_2\kappa_m\mu_0}{4\pi r^2}$, if κ_m is the permeability of the medium in which they are embedded.

CHAPTER VII

ELECTROMAGNETIC INDUCTION AND MAXWELL'S EQUATIONS

The history of electromagnetism has shown as its most conspicuous feature the gradual discovery of interconnections among problems that were at first supposed to be separated. The two oldest fields were electrostatics, and the magnetism of permanent magnets and of ferromagnetic bodies. Early in the nineteenth century Oersted and others demonstrated the magnetic effects of continuous currents, bringing together the study of the electric current, which developed with the discovery of various forms of batteries in the eighteenth century, and the study of magnetism. Faraday, soon after this work, began looking for a converse effect. He reasoned that, if currents could produce magnetism, magnets should be able to produce currents. His first idea was simple, but wrong. He wound two coils of wire together, but insulated from each other, and planned to pass current through one of them, converting it into an electromagnet. He hoped that this magnet, with its lines of induction threading through the other coil, would cause a continuous current to flow through that coil, just as a continuous current produces a continuous magnetic field. His experiment did not show such a current; even though his battery was powerful enough so that his primary coil was heated red-hot, still no current flowed in the secondary. But he was a good enough observer to notice that, though there was no steady current in the secondary, there was nevertheless a transient current when the current was started in the primary, and a transient in the opposite direction when the primary current was interrupted. This suggested to him that the effect he was seeking really existed, but that the induced current for which he was searching was proportional, not to the magnetic flux itself, but to its time rate of change. This law of electromagnetic induction, which bears Faraday's name, is the foundation of the study of electromagnetic theory.

Faraday thought, not in mathematical language, but in terms of lines of force. We have already seen how this concept, leading to the field theory of electrostatics and magnetostatics, allows us to formulate problems involving dielectrics, conductors, and magnetic

materials in a form that is much more powerful than any method based on the concept of action at a distance. In electromagnetic induction as well the idea of lines of force, and of flux, proved to be of the greatest value, the statement of Faraday's law involving the time rate of change of magnetic flux through the coil. The proper mathematical formulation of the methods of lines of force did not come, however, until the work of Maxwell, a number of years later. Maxwell brought together the various laws of electrostatics and magnetostatics which we have already discussed, and correlated them with Faraday's law, expressed in a similar differential or vector form. He was able to see that the system of laws so built up was mathematically inconsistent, and that to make it consistent he had to introduce a certain quantity called the "displacement current," which was in principle susceptible of experimental observation, but which had not been observed at the time. By combining these laws, he set up the equations known as Maxwell's equations, which have been the foundation of the theory of electromagnetism ever since. These equations did not merely describe the phenomena known at Maxwell's time; they predicted the existence of electromagnetic waves, which later proved to be identical with light waves, thus enormously extending the range of the theory. These advances we shall take up in later chapters; we now consider the detailed nature of the law of electromagnetic induction, and its role in electromagnetic theory.

1. The Law of Electromagnetic Induction.—It is found experimentally that, when the flux of magnetic induction **B** through a wire or other conducting circuit changes with time, a current flows in the circuit. The law of electromagnetic induction can be stated most simply, not in terms of the current set up in the circuit, but in terms of the emf leading to the current. By definition, the emf around a circuit equals the total work done, both by electric and magnetic forces and by any other sort of forces, such as those concerned in chemical processes, per unit charge, in carrying a charge around the circuit. Faraday's law states that the emf induced in a circuit equals the negative of the rate of increase of flux of **B** through the circuit. The flux of **B** is by definition the surface integral of the normal component of **B** over a surface whose perimeter is the circuit in question. We may then write Faraday's law in integral form in the following way:

$$\int \mathbf{E} \cdot d\mathbf{s} = -\frac{d}{dt}\int \mathbf{B} \cdot \mathbf{n}\, da. \qquad (1.1)$$

The term on the left is the emf; in the case of electromagnetic induc-

CHAPTER VII

ELECTROMAGNETIC INDUCTION AND MAXWELL'S EQUATIONS

The history of electromagnetism has shown as its most conspicuous feature the gradual discovery of interconnections among problems that were at first supposed to be separated. The two oldest fields were electrostatics, and the magnetism of permanent magnets and of ferromagnetic bodies. Early in the nineteenth century Oersted and others demonstrated the magnetic effects of continuous currents, bringing together the study of the electric current, which developed with the discovery of various forms of batteries in the eighteenth century, and the study of magnetism. Faraday, soon after this work, began looking for a converse effect. He reasoned that, if currents could produce magnetism, magnets should be able to produce currents. His first idea was simple, but wrong. He wound two coils of wire together, but insulated from each other, and planned to pass current through one of them, converting it into an electromagnet. He hoped that this magnet, with its lines of induction threading through the other coil, would cause a continuous current to flow through that coil, just as a continuous current produces a continuous magnetic field. His experiment did not show such a current; even though his battery was powerful enough so that his primary coil was heated red-hot, still no current flowed in the secondary. But he was a good enough observer to notice that, though there was no steady current in the secondary, there was nevertheless a transient current when the current was started in the primary, and a transient in the opposite direction when the primary current was interrupted. This suggested to him that the effect he was seeking really existed, but that the induced current for which he was searching was proportional, not to the magnetic flux itself, but to its time rate of change. This law of electromagnetic induction, which bears Faraday's name, is the foundation of the study of electromagnetic theory.

Faraday thought, not in mathematical language, but in terms of lines of force. We have already seen how this concept, leading to the field theory of electrostatics and magnetostatics, allows us to formulate problems involving dielectrics, conductors, and magnetic

materials in a form that is much more powerful than any method based on the concept of action at a distance. In electromagnetic induction as well the idea of lines of force, and of flux, proved to be of the greatest value, the statement of Faraday's law involving the time rate of change of magnetic flux through the coil. The proper mathematical formulation of the methods of lines of force did not come, however, until the work of Maxwell, a number of years later. Maxwell brought together the various laws of electrostatics and magnetostatics which we have already discussed, and correlated them with Faraday's law, expressed in a similar differential or vector form. He was able to see that the system of laws so built up was mathematically inconsistent, and that to make it consistent he had to introduce a certain quantity called the "displacement current," which was in principle susceptible of experimental observation, but which had not been observed at the time. By combining these laws, he set up the equations known as Maxwell's equations, which have been the foundation of the theory of electromagnetism ever since. These equations did not merely describe the phenomena known at Maxwell's time; they predicted the existence of electromagnetic waves, which later proved to be identical with light waves, thus enormously extending the range of the theory. These advances we shall take up in later chapters; we now consider the detailed nature of the law of electromagnetic induction, and its role in electromagnetic theory.

1. The Law of Electromagnetic Induction.—It is found experimentally that, when the flux of magnetic induction **B** through a wire or other conducting circuit changes with time, a current flows in the circuit. The law of electromagnetic induction can be stated most simply, not in terms of the current set up in the circuit, but in terms of the emf leading to the current. By definition, the emf around a circuit equals the total work done, both by electric and magnetic forces and by any other sort of forces, such as those concerned in chemical processes, per unit charge, in carrying a charge around the circuit. Faraday's law states that the emf induced in a circuit equals the negative of the rate of increase of flux of **B** through the circuit. The flux of **B** is by definition the surface integral of the normal component of **B** over a surface whose perimeter is the circuit in question. We may then write Faraday's law in integral form in the following way:

$$\int \mathbf{E} \cdot d\mathbf{s} = -\frac{d}{dt}\int \mathbf{B} \cdot \mathbf{n}\, da. \tag{1.1}$$

The term on the left is the emf; in the case of electromagnetic induc-

tion, the force acting on the charge is, as we shall soon see, an electric force. The term on the right is the time rate of decrease of magnetic induction through the circuit. The surface integral is to be extended over any surface spanning the circuit. Because $\operatorname{div} \mathbf{B} = 0$, this integral will be the same over any such surface; for if we take two such surfaces, the total flux of **B** out of the volume enclosed by the two surfaces must be zero, so that the flux through each surface must be the same. In (1.1), the emf is to be expressed in volts, and the flux in webers; we note that 1 volt must be equivalent to 1 weber/sec.

It is a familiar fact associated with electromagnetic induction that there are several ways in which the magnetic flux through a circuit can change. First, the circuit itself may move from one part of the magnetic field to another. This is the process used in the ordinary electromagnetic generator, or dynamo. Secondly, the circuit may be stationary, but the magnetic induction may be a function of time, either because it is produced by currents that are varying with time, or because magnetic materials such as permanent magnets are moving. If **B** is changing with time, but the circuit is fixed, we may take the time derivative in (1.1) inside the integral sign. Also we may transform the left side of (1.1) into a surface integral, by using Stokes's theorem. Thus we have

$$\int \operatorname{curl} \mathbf{E} \cdot \mathbf{n}\, da = -\int \left(\frac{\partial \mathbf{B}}{\partial t}\right) \cdot \mathbf{n}\, da.$$

This equation must hold for a circuit of any shape, and with any direction of the normal **n**; thus the integrands on the two sides must be equal, or we have

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{1.2}$$

This equation is the differential form of Faraday's law, and as we shall see later it is one of the fundamental equations of electromagnetic theory, one of Maxwell's equations.

2. Self- and Mutual Induction.—If a current i_1 flows in a circuit, it produces a magnetic field $\mathbf{B}_1$ proportional to i_1 at all points of space. If i_1 changes slowly, the magnetic field is very closely the same at any instant as it would be if the current had a constant value equal to its instantaneous value; such a condition is known as a "quasistationary state." There will then be a time rate of change of $\mathbf{B}_1$ proportional to the time rate of change of i_1. The flux of $\mathbf{B}_1$ through the circuit itself will then vary as i_1 does, so that we shall have an emf in this circuit proportional to di_1/dt, which we may

write as

$$\text{emf}_1 = -L\frac{di_1}{dt},$$

where L is the coefficient of self-induction. Similarly the flux of $\mathbf{B}_1$ through another circuit will vary as i_1 does, so that the emf in a second circuit is

$$\text{emf}_2 = -M\frac{di_1}{dt},$$

where M is the coefficient of mutual induction.

We can easily set up general expressions for the coefficients of self- and mutual induction by use of Eq. (6.5) of Chap. V, expressing the contribution to the vector potential resulting from a current element $i\,\mathbf{ds}$. From the discussion above, L equals the flux of $\mathbf{B}$ through a circuit per unit current flowing in it, and M the flux of $\mathbf{B}$ through a circuit per unit current flowing in another circuit. But remembering that $\mathbf{B} = \text{curl}\,\mathbf{A}$, we have

$$\int \mathbf{B}\cdot\mathbf{n}\,da = \int \text{curl}\,\mathbf{A}\cdot\mathbf{n}\,da = \int \mathbf{A}\cdot\mathbf{ds},$$

in which the last line integral follows from Stokes's theorem. Using Eq. (6.5), Chap. V, which is

$$\mathbf{dA} = \frac{\mu_0}{4\pi}\frac{i\,\mathbf{ds}}{r}, \tag{2.1}$$

and setting $i = 1$, we have

$$L \text{ or } M = \frac{\mu_0}{4\pi}\int\int\frac{\mathbf{ds}\cdot\mathbf{ds}'}{r}. \tag{2.2}$$

Here $\mathbf{ds}$ and $\mathbf{ds}'$ are two elements of length, and r is the distance between them. The integral is a double integral, since each of the elements of length is integrated around the contour. For L, both $\mathbf{ds}$ and $\mathbf{ds}'$ are to be integrated around the single circuit 1; for M, $\mathbf{ds}$ is to be integrated around circuit 1, $\mathbf{ds}'$ around circuit 2. Because of the symmetrical formula, we see that the mutual inductance between two circuits is the same, no matter which of them we regard as the source of the current, and which as the one in which the emf is being induced.

In many special cases, we can make direct calculations of the coefficients of self- and mutual inductance, by simpler methods than the use of (2.2), which in practice usually involves complicated integrations. For, if we can find the magnetic flux resulting from the system of currents in the circuits, we can often integrate these fluxes over the circuits directly. Thus, for instance, consider the

self-inductance per unit length of two parallel wires of radius a, at distance D apart between their centers, shown by Fig. 7, which was used earlier in computing the capacity of the same parallel wires. To find the magnetic field produced by a unit current flowing up through the right-hand wire, down through the left-hand wire, we first consider unit current flowing up through an infinitely thin conductor located at point 2 of the figure, and an equal current flowing down through a linear conductor at point 1. We note an analogy between the electrostatic and magnetic problems: the scalar potential φ in the electrostatic case, for a distribution of charge along these lines, is given by Eq. (3.5) of Chap. I, and the vector potential **A** in the magnetic case for a distribution of current is given by Eq. (6.4) of Chap. V. These formulas are the same, except for two differences: the formula for **A** has μ_0 in the numerator in place of ϵ_0 in the denominator, and the formula for **A** involves current density in place of charge density, and hence results in a vector.

Since every current element is in the same direction in our present problem, **A** in this case will also be in the same direction, normal to the plane of the paper in the figure. If **k** is a unit vector along this direction, and A is the magnitude of **A**, we have $\mathbf{A} = \mathbf{k}A$. To take the curl of **A**, and find **B**, we use the rule of vector analysis that

$$\text{curl}\, f\mathbf{F} = f\,\text{curl}\,\mathbf{F} + \text{grad}\, f \times \mathbf{F},$$

where f is a scalar, **F** a vector. Letting f stand for A, **F** for **k**, and remembering that **F** is then a constant vector, we have

$$\mathbf{B} = \text{curl}\,\mathbf{A} = \text{grad}\, A \times \mathbf{k}.$$

If on the other hand we had had unit charge along the lines instead of unit current, the electric field would have been

$$\mathbf{E} = -\,\text{grad}\,\varphi,$$

where as we have just seen φ equals $A/\epsilon_0\mu_0$. Thus we see that the **B** resulting from unit current, and the **E** resulting from unit charge, are related by the equation

$$\mathbf{B} = \epsilon_0\mu_0(\mathbf{k} \times \mathbf{E}).$$

The magnitudes, in other words, are proportional, and the vectors are at right angles to each other, so that the lines of magnetic force are identical with the electric equipotentials in the problem of Fig. 7. In particular, the circular boundaries of the wires are lines of magnetic force.

Now, to integrate the normal component of **B** over a surface

spanning the circuit, we may integrate over a plane surface including the two lines 1 and 2. The normal component of **B** in this case is just $\epsilon_0\mu_0$ times the tangential component of **E** in the corresponding electrical case. Thus the surface integral of the normal component of **B** over unit length of the transmission line, integrating between the boundaries of the two cylinders, is $\epsilon_0\mu_0$ times the corresponding difference of potential in the electrical case. The inductance is the flux of **B** per unit current; thus it is $\epsilon_0\mu_0$ times the difference of potential per unit charge, or is $\epsilon_0\mu_0$ divided by the electrostatic capacity per unit length. Using Eq. (2.1) of Chap. II for this capacity, we then find that the inductance per unit length is

$$L = \frac{\mu_0}{\pi} \cosh^{-1} \frac{D}{2a}. \tag{2.3}$$

In this calculation, we have neglected magnetic flux through the interior of the wires. This flux would be small in any case, and could almost be neglected. If the wires, instead of being solid, were hollow, so that the current flowed in a thin shell on the surface, and there was no flux inside, we should have just the condition for which (2.3) was correct. The method of calculation we have used here, leading to a simple relationship between problems in electrostatics and magnetism, can often be used, and can be justified in a much more general case than the one we have considered here.

3. The Displacement Current.—In the course of our work, we have derived four fundamental electromagnetic equations, (1.3) of Chap. IV, (3.2) of Chap. V, (3.1) of Chap. VI, and (1.2) of the present chapter, or

$$\begin{aligned} \operatorname{div} \mathbf{D} &= \rho, \\ \operatorname{div} \mathbf{B} &= 0, \\ \operatorname{curl} \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\ \operatorname{curl} \mathbf{H} &= \mathbf{J} \end{aligned} \tag{3.1}$$

These are almost Maxwell's equations, but there is a difficulty with the last of them. We have derived it from Ampère's law, on the basis of steady closed currents, and for this case it is correct. The difficulty occurs when we try to apply the equation to nonstationary cases. Suppose we have a current flowing in an open circuit, as in the discharge of a condenser. The current starts at the positively charged plate, whose charge diminishes as the current flows to the negatively charged plate and annuls the charge there. Thus we can look upon the condenser plates as sources or sinks of current. Now, if we take

the divergence of the last equation, we have

$$\operatorname{div} \operatorname{curl} \mathbf{H} = \operatorname{div} \mathbf{J},$$

and, since the divergence of any curl is zero, we find that div **J** equals zero, which means that the current is always closed and there are no sources or sinks. Thus we are led to a contradiction.

Maxwell concluded from this that the last equation of (3.1) must be incomplete, and that to the term **J** must be added another term, such that the sum of the two had no divergence. We can easily find what this term must be, from the equation of continuity for the flow of current. The divergence of **J** measures the flux outward over the surface of unit volume. If there is current flowing outward, the charge within unit volume must be decreasing, the flux equaling the rate of decrease of charge in unit volume. Thus we have

$$\operatorname{div} \mathbf{J} = -\frac{\partial \rho}{\partial t}. \qquad (3.2)$$

We may rewrite this equation of continuity, using the equation $\operatorname{div} \mathbf{D} = \rho$, and obtaining

$$\operatorname{div}\left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}\right) = 0.$$

In other words, although div **J** is not zero, the divergence of the quantity $\mathbf{J} + \partial\mathbf{D}/\partial t$ is always zero, so that this quantity can mathematically be placed equal to a curl. Maxwell made the assumption that the last equation of (3.1) should properly be replaced by

$$\operatorname{curl} \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \qquad (3.3)$$

The last term $\partial\mathbf{D}/\partial t$ is called the "displacement current," to distinguish it from **J**, the conduction current. By adding this term to Ampère's law, Maxwell assumed that a time rate of change of displacement produced a magnetic field, just as a conduction current does. A test of Maxwell's hypothesis can be made only with very rapidly varying fields, in which the rate of change of **D** with time is so great that the displacement current is large compared with the conduction current, or so that magnetic forces produced by it are comparable with the electric forces due to **E**. We shall see later that these cases are those in which we have electromagnetic waves, whose existence is possible only because of the presence of the displacement current in (3.3). Their existence, then, forms a demonstration of the correctness of Maxwell's hypothesis.

To understand the physical meaning of the displacement current in a simple case, consider the charging of a condenser. Current flows, from one plate through the wire to the other plate. If the current is i, this equals the rate of increase of charge on the plate. Suppose the plates are of area A, separation d, then the value of D between them is $D =$ charge$/A$, and the displacement-current density in the region between plates is current$/A$. Thus the total displacement current is equal to the conduction current in the wire, so that the current is continuous through the circuit.

4. Maxwell's Equations.—We now can write Maxwell's equations,

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \text{div } \mathbf{B} = 0,$$
$$\text{curl } \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, \qquad \text{div } \mathbf{D} = \rho, \tag{4.1}$$

in which we note that the divergence equations follow from the curl equations by taking the divergence, using the equation of continuity (3.2), and integrating with respect to time. These are to be supplemented by the relations (2.1) of Chap. IV, and (2.4) of Chap. VI,

$$\mathbf{D} = \epsilon \mathbf{E}, \qquad \mathbf{B} = \mu \mathbf{H}. \tag{4.2}$$

These are often called the "constitutive equations." If the current density $\mathbf{J}$ obeys Ohm's law, we often include it also in the statement of the constitutive equations. This law is easily stated in differential form by considering a small volume, having length L in the direction of the current, and cross-sectional area A normal to the current. We apply Ohm's law in the form potential difference $= iR$. Here the potential difference is the field $\mathbf{E}$ times the length L of the volume, the current is the area times the current density $\mathbf{J}$, and the resistance is the specific resistance times L/A. Hence we have

$$\mathbf{E}L = A\mathbf{J}\frac{L}{A} \times \text{specific resistance},$$
$$\mathbf{J} = \sigma \mathbf{E}, \tag{4.3}$$

where σ, the specific conductivity, is the reciprocal of the specific resistance. The unit of conductivity is the same as that of J/E; that is, its units are 1/(ohm-meter) = mhos per meter, where 1 mho is defined as the reciprocal of 1 ohm. It should be noted that 1 mho/m is $1/100$ of 1 mho/cm, the usual unit of conductivity. Ohm's law in the form (4.3) can be added to the relations (4.2).

Maxwell's equations, taken together with the constitutive equa-

tions, determine the field, when we are given the charges and currents. To make a complete set of dynamical principles, however, we need two more relations. First is the formula (1.1) of Chap. I,

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}),$$

giving the force acting on a charge moving with a given speed, or the corresponding force acting on the charge and current in unit volume,

$$\mathbf{F} = \rho\mathbf{E} + (\mathbf{J} \times \mathbf{B}). \tag{4.4}$$

Secondly, we must have a law determining the motion of charge in terms of force acting. If the charge is in a metallic conductor obeying Ohm's law, (4.3), that law provides the necessary relation. For electrons and ions moving in empty space, however, as in a vacuum or discharge tube, we must use Newton's law, that the force equals the mass times the acceleration, or the time rate of change of the momentum. With such a law, we find the field from the charge, the force from the field, and the motion from the force, obtaining therefore a complete system of dynamics.

Maxwell's equations, the constitutive equations, and the force equation, as just written, form the foundation of the whole of electromagnetic theory, and as far as is known are exactly correct, aside from the corrections resulting from the quantum theory, which in a sense change the whole formulation of the theory. They allow the derivation of the electromagnetic theory of light, and of electromagnetic waves in general, which we shall shortly take up. They hold even for particles moving nearly with the velocity of light, for which the theory of relativity must be used in discussing the mechanical part of the motion, but for which Maxwell's methods are still correct for finding the forces. In the older developments of electrical engineering, the so-called "lumped-circuit theory," it was possible to operate with the limiting case of slow variation with time, in which the displacement current could be neglected. The newer developments of microwaves and distributed constants, however, operate with the theory of electromagnetic waves, for which the displacement current is essential, and Maxwell's equations as formulated in this section, rather than the quasistationary form of them as given in (3.1), must be used.

5. The Vector and Scalar Potentials.—We observe that, if $\mathbf{B}$ depends on time, curl $\mathbf{E} \neq 0$, so that there is no potential for $\mathbf{E}$. The ordinary electrical potential is thus confined to static problems. Further, if $\mathbf{J}$ or $\partial\mathbf{D}/\partial t \neq 0$, there is no potential for $\mathbf{B}$. We have

seen in Chap. V how a potential can be introduced for **B**: one uses a vector potential **A**, possible because div **B** = 0. That is, we let

$$\mathbf{B} = \text{curl } \mathbf{A}. \tag{5.1}$$

We can do this even in the general case. And it proves that we can use a scalar potential φ, reducing to the electrostatic potential in the case of a steady state, but different in other cases, by a special device. The relation that proves to be satisfied is that

$$\mathbf{E} = -\text{ grad } \varphi - \frac{\partial \mathbf{A}}{\partial t}, \tag{5.2}$$

reducing to the familiar $\mathbf{E} = -\text{ grad } \varphi$ when the vector potential is independent of time. To verify these statements, we substitute the expressions for **E** and **B** in Maxwell's equations, and see if they can be satisfied by proper choice of **A** and φ. First we note that

$$\text{div } \mathbf{B} = \text{div curl } \mathbf{A}$$

is automatically zero, since the divergence of any curl is zero. Similarly, remembering that the curl of any gradient is zero, we find that the equation curl $\mathbf{E} = -\partial \mathbf{B}/\partial t$ is automatically satisfied.

Next we must consider the other two equations, in **D** and **H**. To find these quantities in terms of **E** and **B**, we must use the constitutive equations. The relations are simple only if ϵ and μ are constants independent of position, and we consider only that case. From the equation curl $\mathbf{H} - \partial \mathbf{D}/\partial t = \mathbf{J}$, we have, using the relation of vector analysis that

$$\text{curl curl } \mathbf{A} = \text{grad div } \mathbf{A} - \nabla^2 \mathbf{A},$$

the result

$$\nabla^2 \mathbf{A} - \epsilon\mu \frac{\partial^2 \mathbf{A}}{\partial t^2} - \text{grad}\left(\text{div } \mathbf{A} + \epsilon\mu \frac{\partial \varphi}{\partial t}\right) = -\mu \mathbf{J}.$$

Similarly, from the equation div $\mathbf{D} = \rho$, with a little manipulation, we find

$$\nabla^2 \varphi - \epsilon\mu \frac{\partial^2 \varphi}{\partial t^2} + \frac{\partial}{\partial t}\left(\text{div } \mathbf{A} + \epsilon\mu \frac{\partial \varphi}{\partial t}\right) = -\frac{\rho}{\epsilon}.$$

Now let us choose **A** and φ subject to the condition that

$$\text{div } \mathbf{A} + \epsilon\mu \frac{\partial \varphi}{\partial t} = 0. \tag{5.3}$$

Since div **A** is so far arbitrary, we can do this. Then the equations

for the potentials become

$$\nabla^2 \mathbf{A} - \epsilon\mu \frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu \mathbf{J},$$
$$\nabla^2 \varphi - \epsilon\mu \frac{\partial^2 \varphi}{\partial t^2} = -\frac{\rho}{\epsilon}. \tag{5.4}$$

If **A** and φ satisfy these equations, then, as we stated before, the fields determined from them by (5.1) and (5.2) satisfy Maxwell's equations. The equations for the potentials are of the form called "d'Alembert's equation," and as can be seen are extensions of Poisson's equation, obtained by adding the time derivatives. We observe that, in regions where there is no charge and current density, the potentials satisfy the wave equation, which is the homogeneous equation obtained by setting the right side of d'Alembert's equation equal to zero. We shall show in the next chapter that this means that φ and **A** are given by functions representing waves traveling with the velocity $1/\sqrt{\epsilon\mu}$, and that the same thing is true of the fields **E** and **H**.

In (5.4) we have set up equations for A and φ in terms of prescribed values of the charge and current density. Sometimes, however, we wish to assume that the current density obeys Ohm's law, and that the charge density is zero, as we should have in the interior of a conductor. In that case, assuming (4.3) for **J**, and proceeding in a similar manner, we find that in place of (5.3) we should assume

$$\operatorname{div} \mathbf{A} + \sigma\mu\varphi + \epsilon\mu \frac{\partial \varphi}{\partial t} = 0,$$

and in place of the equations (5.4) we have

$$\nabla^2 \mathbf{A} - \sigma\mu \frac{\partial \mathbf{A}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{A}}{\partial t^2} = 0,$$
$$\nabla^2 \varphi - \sigma\mu \frac{\partial \varphi}{\partial t} - \epsilon\mu \frac{\partial^2 \varphi}{\partial t^2} = 0. \tag{5.5}$$

We shall find that these equations, involving first as well as second time derivatives, represent damped or attenuated waves, as we should expect in an absorbing medium, such as a metallic conductor.

Problems

1. Two parallel conducting strips, of width w, distance of separation d (where $d \ll w$), indefinitely long, carry current in opposite directions, so as to form the two conductors of an electric circuit. Find the self-inductance of such a circuit per unit length.

2. Two concentric thin-walled hollow conducting cylinders, of radii r_1 and r_2, carry current in opposite directions. Find the self-inductance per unit length. Find how the result is changed if the inner conductor is a solid conducting rod, carrying current uniformly distributed through its interior.

3. Find the self-inductance of a circuit consisting of a hollow cylindrical conductor of radius r, parallel to an infinite conducting sheet that forms the return path of the current. (*Suggestion:* Use the method of images.)

4. A magnetic field points along the z axis, and its magnitude is proportional to time, and independent of position. Find the vector potential. Assuming that the scalar potential is zero, find the induced electric field. Prove by direct integration, using a circular path, that the law of induction holds.

5. Describe the magnetic field between the plates of a condenser while it is being charged up.

6. Starting from the induction law, show that the line integral of $\left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right)$ around a closed path is zero, where $\mathbf{A}$ is the vector potential. From this show that the curl of the above vector vanishes, and hence that $\mathbf{E} = -\operatorname{grad} \varphi - \frac{\partial \mathbf{A}}{\partial t}$, where φ is the scalar potential.

7. Derive the differential equations satisfied by $\mathbf{A}$ and φ for quasistationary processes.

8. Starting with the equation of continuity, and assuming Ohm's law, show that the charge density in a conductor obeys the equation $\frac{\sigma}{\epsilon}\rho + \frac{\partial \rho}{\partial t} = 0$. Show that any existing charge distribution within a conductor will be damped off exponentially, and find the time required for it to be reduced to $1/e$ of its initial value. Insert numerical values for copper, and find the value of this time, the relaxation time. Similarly find the relaxation time for a good insulator, such as quartz.

9. A long solenoid of n turns per unit length carries an alternating current of angular frequency ω and peak value I. If a nonmagnetic rod of conductivity σ just filling the solenoid is placed inside it, find the current distribution, the magnetic field, and the heating loss per unit length inside the copper rod. (The heat loss per unit volume per second equals $\mathbf{J} \cdot \mathbf{E}$.) Discuss the behavior of your solution for the limiting cases of large and small conductivities. How would your results be altered if the rod were coaxial with the solenoid but of smaller radius?

10. The outer conductor of a coaxial line as in Prob. 2 is a thin-walled cylinder of radius a and thickness t. Show that the contribution to the inductance per unit length of the line arising from the magnetic field in this conductor is given very nearly by

$$\frac{\mu_0}{6\pi}\frac{t}{a}, \qquad \text{where } \frac{t}{a} \ll 1.$$

11. Show that, in a nondissipative medium if the divergence of the vector potential is set equal to zero, instead of satisfying Eq. (5.3), the scalar potential satisfies Poisson's equation with ρ, the density of electric charge, a function of coordinates and time. Find the differential equation satisfied by the vector potential under these conditions.

CHAPTER VIII

ELECTROMAGNETIC WAVES AND ENERGY FLOW

The first and most conspicuous success of Maxwell's theory was its prediction of the existence of electromagnetic waves, whose velocity of propagation equaled the experimentally known velocity of light. It was immediately clear that light must be a form of electromagnetic radiation, of short wave length. It was a number of years later that Hertz demonstrated experimentally the existence of electromagnetic waves of longer wave length, but when they were found, they proved, like light, to satisfy Maxwell's equations. In the present chapter we shall discuss the simplest forms of waves, plane waves; we come to more complicated waves, such as spherical waves, in a later chapter. We also consider an aspect of the electromagnetic field that we have so far passed over: the energy associated with it. We have so far treated problems by considering only the forces acting, rather than the work done and the resulting energy. With radiation, however, it is so obvious that the field carries energy that without a consideration of the energy flow we can hardly form a correct picture of the phenomena. We therefore take up, not only the density of electric and magnetic energy in the field, which ties in with well-known facts regarding the energy in condensers and inductances, but also Poynting's theorem, a deduction from Maxwell's equations which provides a simple way of finding the flow of energy in any electromagnetic field.

1. Plane Waves and Maxwell's Equations.—Let us consider the problem of solving Maxwell's equations in a uniform material having dielectric constant $\kappa_e = \epsilon/\epsilon_0$, magnetic permeability $\kappa_m = \mu/\mu_0$, conductivity σ, but not any charge, or any current other than that determined by Ohm's law. Maxwell's equations in this case are

$$\text{curl } \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \text{div } \mathbf{H} = 0,$$

$$\text{curl } \mathbf{H} = \sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \text{div } \mathbf{E} = 0. \tag{1.1}$$

We take the curl of the first equation and substitute from the second

for curl **H**, obtaining

$$\text{curl curl } \mathbf{E} = -\sigma\mu \frac{\partial \mathbf{E}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2}.$$

Similarly, we take the curl of the second and substitute from the first, obtaining

$$\text{curl curl } \mathbf{H} = -\sigma\mu \frac{\partial \mathbf{H}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2}.$$

Using the equation of vector analysis,

$$\text{curl curl } \mathbf{A} = \text{grad div } \mathbf{A} - \nabla^2 \mathbf{A},$$

where **A** is an arbitrary vector function, and using the equations div $\mathbf{E} = 0$, div $\mathbf{H} = 0$, these equations become

$$\nabla^2 \mathbf{E} - \sigma\mu \frac{\partial \mathbf{E}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0,$$
$$\nabla^2 \mathbf{H} - \sigma\mu \frac{\partial \mathbf{H}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2} = 0. \tag{1.2}$$

Thus **E** and **H** satisfy the same wave equation that we have already found in Eq. (5.5), Chap. VII, for the potentials **A** and φ. We can of course get the case of a nonconducting medium by setting $\sigma = 0$, in which case the middle term of the equation is omitted. The equations (1.2) are vector equations, which means that each of the six components of **E** and **H** separately satisfies the same scalar wave equation.

With a wave equation of the form of

$$\nabla^2 u - \sigma\mu \frac{\partial u}{\partial t} - \epsilon\mu \frac{\partial^2 u}{\partial t^2} = 0, \tag{1.3}$$

where u, a scalar, can stand for one of the components of **E, H, A,** or φ, we can find a great variety of solutions. In fact, a good deal of the remainder of our study will be devoted to different solutions of this equation. We shall start with the simplest, and in many ways the most important, of these solutions, the plane wave, understanding that it is a special case, rather than a general solution. We shall assume that u is a function of z, one of the space coordinates, and of t, only, being independent of x and y, or constant on planes of constant z. Furthermore, we shall assume that u varies with both z and t in an exponential manner. We shall write our assumptions in the form

$$u = u_0 e^{j\omega t - \gamma z}, \tag{1.4}$$

where $j = \sqrt{-1}$, and u_0, ω, and γ are constants. We shall describe the physical significance of these constants in a moment. Substituting (1.4) in the wave equation (1.3), we find

$$\gamma = \pm \sqrt{\sigma\mu j\omega - \epsilon\mu\omega^2} = \pm j\omega\sqrt{\left(\epsilon - \frac{j\sigma}{\omega}\right)\mu}. \tag{1.5}$$

In other words, if γ is given by either of the values (1.5), the expression (1.4) is a solution of the wave equation.

From the form of (1.5) we see that γ is in general complex. We may write it

$$\gamma = \alpha + j\beta, \tag{1.6}$$

where α and β are real and imaginary parts, respectively. We note that the quantity whose square root is being taken in (1.5) lies in the second quadrant in the complex plane, its real part being negative and its imaginary part positive; thus the positive square root lies in the first quadrant, so that α and β are both positive. If in particular the medium is nonconducting, so that σ is zero, we have $\gamma = j\beta$, where β is positive. We may now rewrite the solution (1.4) in the form

$$u = u_0 e^{\mp\alpha z} e^{j(\omega t \mp \beta z)}. \tag{1.7}$$

As in the theory of oscillating circuits, for physical purposes we use the real part of this complex expression to represent the real value of u. Thus (1.7) represents a disturbance whose value at a given point of space varies sinusoidally with time, with angular frequency ω, or frequency $f = \omega/2\pi$. It also varies sinusoidally with z, in case $\alpha = 0$, with wave length $\lambda = 2\pi/\beta$, so that, when z increases by a wave length, the disturbance reverts to its initial value. A given wave crest moves with a velocity v, given by

$$v = \frac{\omega}{\beta} = \lambda f,$$

the velocity being positive, or in the direction of increasing z, for the upper sign in (1.7), and negative for the lower sign. The factor $e^{\mp\alpha z}$ represents a falling off of intensity, or damping, in the direction of propagation of the wave, such that the amplitude falls to $1/e$ of its value in a distance $1/\alpha$. We see that the disturbance in a conducting medium is then a damped plane wave, and in a nonconducting dielectric it is an undamped plane wave.

For an undamped wave in a nonconducting medium, the velocity of propagation takes a simple form. In that case, setting $\sigma = 0$, we

have from (1.5) and (1.6),

$$\beta = \omega \sqrt{\epsilon\mu}, \qquad v = \frac{\omega}{\beta} = \frac{1}{\sqrt{\epsilon\mu}}. \tag{1.8}$$

If we set

$$\epsilon = \kappa_e\epsilon_0, \qquad \mu = \kappa_m\mu_0,$$

following Eqs. (2.1) of Chap. IV, and (2.4) of Chap. VI, we find

$$v = \frac{c}{n}, \qquad c = \frac{1}{\sqrt{\epsilon_0\mu_0}} = 3.00 \times 10^8 \text{ m/sec},$$
$$n = \sqrt{\kappa_e\kappa_m}. \tag{1.9}$$

That is to say, the velocity of the electromagnetic waves in empty space is 3×10^8 m/sec, which is known to be the velocity of light. We have already mentioned that it was this fact which originally led to the conviction that Maxwell's equations furnished a theory of light, and that Hertz's discovery of electromagnetic waves has led to a series of waves of this type which can exist for frequencies over apparently an unlimited range. We now are familiar with waves ranging from ordinary radio waves at frequencies of a few kilocycles per second, short-wave radio at a few megacycles per second, microwaves with thousands or tens of thousands of megacycles per second, or wave lengths down to the order of a centimeter, through the infrared with wave lengths from fractions of a millimeter down to the visible spectrum with wave lengths of a few thousand angstroms, through the X-ray region with wave lengths down to a few hundredths of an angstrom, and from there through the gamma rays down to wave lengths whose lower limit is so far not known. All these waves satisfy the wave equation, all are essentially the same type, and in our subsequent treatment we shall consider them all.

The index of refraction n, defined in (1.9) as the ratio of the velocity of light in free space to the velocity in a material medium, is given in terms of the dielectric constant and magnetic permeability. Experimentally, this relation holds very accurately for radio waves, and in some materials for the infrared region, but it is almost never correct in the visible region or for shorter wave lengths. The reason will be taken up in a later chapter. It is essentially the fact that the dielectric constant is a function of frequency, rather than a true constant. The charges inside an atom or molecule, which become polarized to lead to the dielectric phenomena, as we have described in Chap. IV, have a certain mass and consequent inertia. For this reason they do not polarize instantaneously, but lag to some extent

behind the field that polarizes them. The treatment of Chap. IV holds only for the steady state, or for the static dielectric constant. In a later chapter we shall take up the time variation, showing that we obtain a correct picture of the variation of the index of refraction with frequency, or of the phenomenon of dispersion.

2. The Relation between E and H in a Plane Wave.—If we substitute expressions of the form (1.4), with constant vectors $\mathbf{E}_0$ and $\mathbf{H}_0$ rather than the scalar u_0 multiplying the exponential, into Maxwell's equations (1.1), we find the following equations from the various components of Maxwell's equations:

$$\begin{aligned}
\gamma E_y &= -\mu j\omega H_x\\
-\gamma E_x &= -\mu j\omega H_y\\
0 &= -\mu j\omega H_z\\
-\gamma H_z &= 0\\
\gamma H_y &= (\sigma + \epsilon j\omega)E_x\\
-\gamma H_x &= (\sigma + \epsilon j\omega)E_y\\
0 &= (\sigma + \epsilon j\omega)E_z\\
-\gamma E_z &= 0.
\end{aligned}$$

We notice in the first place that E_z and H_z are zero; that is, there is no component of **E** or **H** in the direction of propagation, or the wave is transverse. The remaining equations can be written in the form

$$\frac{E_x}{H_y} = -\frac{E_y}{H_x} = Z_0 = \frac{\mu j\omega}{\gamma} = \frac{\gamma}{\sigma + \epsilon j\omega} = \pm\sqrt{\frac{\mu}{\epsilon - j\sigma/\omega}}, \tag{2.1}$$

in which the various forms of the constant Z_0 are related through (1.5). These express the fact that **E** and **H** are at right angles to each other, as well as being at right angles to the direction of propagation, and that the ratio of the magnitude of **E** to the magnitude of **H** is the constant Z_0. We may write this expression in a vector form, if **k** is unit vector along the z axis; we have

$$\mathbf{k} \times \mathbf{E} = Z_0\mathbf{H}, \qquad \mathbf{k} \times \mathbf{H} = -\frac{\mathbf{E}}{Z_0}. \tag{2.2}$$

In the case of empty space, we have

$$Z_0 = \pm\sqrt{\frac{\mu_0}{\epsilon_0}} = \pm\sqrt{\frac{4\pi \times 10^{-7}}{8.85 \times 10^{-12}}} = \pm 376.6 \text{ ohms}, \tag{2.3}$$

where the units of Z_0 are most easily seen from the fact that it measures a ratio of **E,** in volts per meter, to **H,** in ampere-turns per meter, and therefore must equal volts/amperes or ohms. Because the units of

E/H are the same as those of impedance, the value of Z_0 is often referred to as the "wave impedance" of the medium, and the particular value (2.3) is the wave impedance of empty space, which in that particular case is a pure resistance, though we see from (2.1) that in a conducting medium there is a reactive as well as a resistive component.

3. Electric and Magnetic Energy Density.—In our treatment of electrostatics and magnetostatics we worked entirely with forces rather than introducing the concept of energy. In Sec. 3, Chap. I, however, we introduced the electrostatic potential, from which we can see at once that the energy of two charges q_1 and q_2, at a distance r_{12} from each other, is

$$V_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}}. \tag{3.1}$$

The total electrostatic energy of a system of charges is then the sum of expressions like (3.1) over all pairs of charges. A sum over pairs of terms like V_{ij} can also be written as half the double sum of the same quantities over the indices i and j separately (omitting of course the case $i = j$); the double sum includes each pair of indices twice, as for instance the terms V_{12} (for $i = 1, j = 2$) and V_{21}, which equals it (for $i = 2, j = 1$), and to compensate this we must have a factor $\frac{1}{2}$. Thus we have for the total potential energy

$$V = \frac{1}{2} \sum_i q_i \sum_{j \neq i} \frac{q_j}{4\pi\epsilon_0 r_{ij}}. \tag{3.2}$$

But the summation over j, in the expression above, is just the potential φ_i at the location of the charge q_i. Thus we may rewrite (3.2) in the form

$$V = \tfrac{1}{2} \sum_i q_i \varphi_i.$$

If the charges are continuously distributed, we may replace the summation over charges by an integration, replacing q_i by $\rho\, dv$, the charge in the volume dv. Thus we have

$$V = \tfrac{1}{2} \int \rho\varphi\, dv. \tag{3.3}$$

This is the standard expression for the energy of a system of charges, in terms of the charge density and the potential.

The expression (3.3) can be rewritten in terms of the field, rather than of the charge and potential. We have, for the electrostatic case,

$$\mathbf{E} = -\operatorname{grad} \varphi, \qquad \operatorname{div} \mathbf{D} = -\operatorname{div} (\epsilon \operatorname{grad} \varphi) = \rho,$$

which is the form that Poisson's equation would take in general where ϵ may be a function of position. Using this result, (3.3) takes the form

$$V = -\tfrac{1}{2}\int \varphi \operatorname{div} (\epsilon \operatorname{grad} \varphi)\, dv. \tag{3.4}$$

But by simple vector analysis we have

$$\operatorname{div} (\epsilon\varphi \operatorname{grad} \varphi) = \varphi \operatorname{div} (\epsilon \operatorname{grad} \varphi) + \epsilon(\operatorname{grad} \varphi)^2.$$

Substituting in (3.4), the energy V becomes

$$V = \tfrac{1}{2}\int \epsilon E^2\, dv + \tfrac{1}{2}\int \epsilon\varphi \mathbf{E} \cdot \mathbf{n}\, da.$$

If we integrate over such a large volume that φ and $\mathbf{E}$ are negligible around the surface, the surface integral vanishes, leaving us with

$$V = \tfrac{1}{2}\int \epsilon E^2\, dv. \tag{3.5}$$

This is the expression for the electrostatic energy in terms of the field. Its interpretation is that we may imagine the energy to be localized throughout the field.

For instance, suppose we take a condenser of capacity C, and let its charge at a given moment be q. Assume that we are charging the condenser, and that we wish to know how much work we shall have to do on it to charge it. To take a small additional charge dq around the circuit, against the difference of potential q/C, will require an amount of work $(q/C)\, dq$. Thus the whole work done in setting up a charge Q is

$$\int_0^Q \frac{q}{C}\, dq = \frac{1}{2}\frac{Q^2}{C}.$$

But if the condenser consists of two plates of area A, distance of separation d, filled with a dielectric of dielectric constant ϵ/ϵ_0, the value of D inside the condenser is Q/A, and the value of E is $(Q/\epsilon)A$. Since the capacity C is $\epsilon A/d$, we then have

$$\frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}\frac{(\epsilon AE)^2}{\epsilon A}\, d = \frac{1}{2}\,(\epsilon E^2)Ad.$$

But, since Ad is the volume of the region in which the field is different from zero, this agrees with the expression (3.5). Thus we see in this example how we can interpret (3.5) as meaning that energy is located throughout the field, with a density $\tfrac{1}{2}\epsilon E^2$.

In a similar way we can consider the magnetic energy to be localized in space. A complete treatment of the magnetic case is considerably more complicated than in the electrical case, and we shall give only a partial discussion; the reader who is interested in a more

complete treatment may find one in J. A. Stratton, *Electromagnetic Theory*, Secs. 2.14 to 2.18 (McGraw-Hill Book Company, Inc., New York, 1941). In the first place, let us consider a circuit of self-inductance L, carrying an instantaneous current i. If i increases, the emf is $-L\,di/dt$, and the work that an external emf must do to bring about an increase di of the current is $Li(di/dt)\,dt = Li\,di$. Thus the total work that must be done to build up a current I in the circuit is

$$T = \int_0^I Li\,di = \tfrac{1}{2}LI^2.$$

Using Eqs. (2.1) and (2.2) of Chap. VII, this may be written

$$T = \tfrac{1}{2}\int I\,\mathbf{ds}\cdot\mathbf{A}. \tag{3.6}$$

In this expression, the integration is around the circuit, I is the current flowing in the element $\mathbf{ds}$, and $\mathbf{A}$ is the vector potential at that point. Using an argument like that above, we can also include the mutual effect when many currents are flowing; the only difference is that $\mathbf{A}$ in (3.6) must be the resultant vector potential not only of the one circuit but of all circuits. If the current densities are distributed throughout space, the current density being $\mathbf{J}$, we replace an integral about a circuit by a volume integral, and have

$$T = \tfrac{1}{2}\int \mathbf{J}\cdot\mathbf{A}\,dv. \tag{3.7}$$

The analogy of this expression to (3.3) is obvious, and it shows us for the first time that the vector potential has much the same relation to the energy of currents as the scalar potential has to the energy of charges.

We may now carry out a transformation, as with the electrostatic case, so as to write this energy T as an integration of the magnetic field rather than of the current density and vector potential. We have

$$\operatorname{div}(\mathbf{H}\times\mathbf{A}) = \mathbf{A}\cdot\operatorname{curl}\mathbf{H} - \mathbf{H}\cdot\operatorname{curl}\mathbf{A} = \mathbf{J}\cdot\mathbf{A} - \mathbf{H}\cdot\mathbf{B},$$

in which we have used Ampère's law $\mathbf{J} = \operatorname{curl}\mathbf{H}$ in the form that holds for the stationary case. Using this result we may rewrite (3.7) in a form involving a volume integral of $\mathbf{H}\cdot\mathbf{B}$, and a surface integral of the normal component of $\mathbf{H}\times\mathbf{A}$. As before, if we integrate over all space, we may assume that the fields are so small over the infinitely distant outer boundary that the surface integral vanishes, and we have finally

$$T = \tfrac{1}{2}\int \mathbf{H}\cdot\mathbf{B}\,dv = \tfrac{1}{2}\int \mu H^2\,dv, \tag{3.8}$$

in which the second form holds when $\mathbf{B} = \mu\mathbf{H}$. In case this is not so,

and we have a ferromagnetic medium, the situation is much more complicated, and we shall not consider that case. The expression (3.8) is clearly analogous to (3.5) for the electrostatic energy. In the special case of the solenoid we may at once verify the result (3.8), as we did in the condenser for the electrostatic case. Let us take a solenoid of N turns, length d, area A, in a medium of permeability μ/μ_0. If a current i flows in it, the magnetic field inside will be $(N/d)I$, and the induction B will be $(N/d)\mu i$. The flux of B through the N turns will then be NA times B, or $(\mu N^2A/d)i$, so that the self-inductance will be $L = \mu N^2A/d$. The magnetic energy T is then

$$\frac{1}{2} Li^2 = \frac{1}{2}\frac{\mu N^2 A}{d}\left(\frac{Hd}{N}\right)^2 = \frac{1}{2}(\mu H^2)Ad,$$

which agrees with (3.8), and fits in with the assumption that the density of magnetic energy is $\frac{1}{2}\mu H^2$.

The expression (3.5) for the electrostatic energy is convenient for solving certain types of problems, because we can find the force on a charged body by stating that the work done when it is displaced is the change in the potential energy V, assuming the charges to remain fixed. For example, we can find the force acting to pull a piece of dielectric from one part of a field to another, by solving the problem of the field distribution as a function of the location of the piece of dielectric, computing the energy as a function of position, and seeing how it changes when the dielectric moves. The corresponding magnetic problem is much more involved, however, and is discussed by Stratton, in the reference given above.

If we take a circuit carrying a certain current, and displace it in a magnetic field, keeping the field and the current constant, the work that we must do proves not to be dT, as we should expect at first sight, but $-dT$. On the other hand, in moving the circuit through the field, the flux through it has changed, and this has resulted in an emf in the circuit. This emf would have resulted in a change of the current through the circuit, which is contrary to our hypothesis that the current is unchanged. Thus we must have been exerting a counter-emf to keep the current constant, and we find that the work that our counter-emf has had to do on the current during the displacement is just dT. Thus the work that our counter-emf does just balances the work that we get out of the system by the displacement of the circuit, and the net work done is zero. This is consistent with an observation that we might have made from Eq. (1.1), Chap. I, in which we state that the force exerted on a charge q moving with a

velocity $\mathbf{v}$, in a magnetic induction $\mathbf{B}$, is $q\mathbf{v} \times \mathbf{B}$. This force is at right angles to the velocity of the charge, and therefore the work done by it is zero.

On the other hand, we are assuming that the field stays constant while the circuit is being moved. Actually this field must be produced by other currents in other circuits. As the circuit is moved, the flux produced by it at these other circuits changes, and emf's are induced in them. To keep the field constant, and hence the currents in those circuits constant, we must have been exerting counter-emf's in those other circuits as well, and the work we have done to maintain those counter-emf's proves again to be dT. Thus in the process we have a net amount of work dT, made up of $-dT$ as the direct work involved in moving the circuit, dT in keeping the current constant in the circuit we are moving, and another dT in keeping the current constant in the circuits that produce the external magnetic field. Unless careful account is taken of these various terms in the energy, whose existence we have merely mentioned without proof, there is great danger of making a mistake in sign when finding the mechanical forces acting on a circuit from the magnetic energy.

These relations are similar to those found in mechanics in certain cases, in which the kinetic energy depends on the coordinates. In such a case we find that the force is given by the derivative, not of the energy, but of the difference between the potential and kinetic energies, with respect to the coordinates. This difference is essentially the Lagrangian function. In a similar way here the force is given by the derivative, not of the energy $V + T$, but of the difference $V - T$, which plays the part of a Lagrangian function, the magnetic energy T playing the part of a kinetic energy. This is not unnatural; for the electrostatic energy V results from the positions of charges, but the magnetic energy T results from their motions, the current i in the expression $\frac{1}{2}Li^2$ being proportional to the velocities of the charges.

4. Poynting's Theorem and Poynting's Vector.—In the preceding sections we have seen that in the static case we have an energy density $\frac{1}{2}(\epsilon E^2 + \mu H^2)$ of electric and magnetic energy. We shall now show that this formula still can be used when the fields are changing with time. To show this we shall first prove a mathematical theorem, Poynting's theorem, and shall then show its interpretation. By a vector identity we have

$$\begin{aligned}\operatorname{div}(\mathbf{E} \times \mathbf{H}) &= \mathbf{H} \cdot \operatorname{curl} \mathbf{E} - \mathbf{E} \cdot \operatorname{curl} \mathbf{H} \\ &= -\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} - \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} - \mathbf{E} \cdot \mathbf{J},\end{aligned}$$

where in the last expression we have used Maxwell's equations. If we assume $\mathbf{D} = \epsilon\mathbf{E}$, $\mathbf{B} = \mu\mathbf{H}$, we have

$$\operatorname{div}(\mathbf{E} \times \mathbf{H}) + \frac{\partial}{\partial t}\frac{1}{2}(\epsilon E^2 + \mu H^2) = -\mathbf{E} \cdot \mathbf{J}. \tag{4.1}$$

This equation reminds us of an ordinary equation of continuity, which states that the divergence of the flux of any quantity, plus the rate at which the density of the quantity increases with time, equals the rate at which the quantity is produced. In other words, applying the equation to a small volume, it states that the rate at which the quantity increases within the volume equals the rate at which it is produced within the volume, minus the rate at which it flows out over the surface. The quantity $-\mathbf{E} \cdot \mathbf{J}$ represents the rate at which energy is produced (that is, $\mathbf{E} \cdot \mathbf{J}$ represents the rate at which energy is lost) per unit volume on account of ordinary joulean or resistance heating. Thus the quantity for which (4.1) forms the equation of continuity is the energy. We are then justified in interpreting the quantity $\frac{1}{2}(\epsilon E^2 + \mu H^2)$ as the energy density, extending the results of the preceding sections to time varying fields. At the same time, we must interpret the vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H},$$

which is known as "Poynting's vector," as the flux of energy, the amount of energy crossing unit area perpendicular to the vector, per unit time.

The conception of the energy of the electromagnetic field as residing in the medium is a very fundamental one, and has had great influence in the development of the theory. Maxwell and his immediate followers thought of the medium as resembling an elastic solid, the electrical energy representing the potential energy of strain of the medium, the magnetic energy the kinetic energy of motion. Though such a mechanical view is no longer held, still the energy is regarded as being localized in space, and as traveling in the manner indicated by Poynting's vector. Thus, in a light wave, we shall show that there is a certain energy per unit volume, proportional to the square of the amplitude (E or H). This energy travels along, and Poynting's vector is the vector that measures the rate of flow, or the intensity of the wave. We have already seen that $\mathbf{E}$ and $\mathbf{H}$ in a plane wave are at right angles to each other, and at right angles to the direction of flow; thus $\mathbf{E} \times \mathbf{H}$ must be along the direction of flow. In more complicated waves as well, Poynting's vector points along the direction of the flow of radiation. If, for example, we have a source of

light, and we wish to find at what rate it is emitting energy, we surround it by a closed surface, and integrate the normal component of Poynting's vector over the surface. The whole conception of energy being transported in the medium is fundamental to the electromagnetic theory of light.

When we come to charges and currents, however, it is a little harder to see the significance of the energy in the medium. For example, in a circuit consisting of a battery, and a wire connecting the plates, Poynting's vector indicates that the energy flows out of the battery through the space surrounding the wire, and finally flows into the wire at the point where it will be transformed into heat. This seems to have small physical significance. With a moving charge, the situation is somewhat more reasonable. A simple model of an electron, which was supposed before the quantum theory to represent its actual structure, was a sphere of radius R, on the surface of which the charge is distributed. Then the field E will be $e/4\pi\epsilon_0 r^2$ at any point outside the sphere, where e is the charge. The total electrical energy is the volume integral of $(e^2/32\pi^2\epsilon_0)(1/r^4)$ over all space outside the sphere, or

$$\frac{e^2}{32\pi^2\epsilon_0}\int_R^\infty \frac{4\pi r^2}{r^4}\,dr = \frac{e^2}{8\pi\epsilon_0 R}.$$

In the classical theory of the electron, which we have mentioned, it is this quantity which is interpreted as being the actual constitutive energy of the electron, though a correction must be made of an additional energy of a nonelectromagnetic nature that is required to keep the sphere in equilibrium. Neglecting this correction, we can compute the mass of the electron. For a relation of Einstein's relativity theory says that a given energy has a mass, given by the relation energy $= mc^2$. Hence $mc^2 = e^2/8\pi\epsilon_0 R$. Solving for the radius, we have $R = e^2/8\pi\epsilon_0 mc^2$, a familiar formula for the radius of the electron. The more correct formula, inserting the correction we omitted, differs only by a small factor. Using the values $e = 1.60 \times 10^{-19}$ coulomb, $m = 9.1 \times 10^{-31}$ kg, $c = 3.00 \times 10^8$ m/sec, $\epsilon_0 = 8.85 \times 10^{-12}$ farad/m, we have $R = 1.42 \times 10^{-15}$ m $= 1.42 \times 10^{-13}$ cm. Now, if this electron moves, it will produce a magnetic field, as a current would, and hence will have a certain magnetic energy. Since the magnetic field is proportional to the velocity (or the current), the magnetic energy is proportional to the square of the velocity. This can be shown to be the kinetic energy. Further, there will be a Poynting vector, pointing in general in the direction of travel of the electron, and representing the flow of energy associated with the electron. All

these relations prove on closer examination to be more complicated than they seem at first sight, but they are suggestive in pointing one possible way to an eventual theory of the structure of the electron, which even the present quantum theory is unable to supply completely.

5. Power Flow and Sinusoidal Time Variation.—We shall often want to find Poynting's vector, and the energy density, in cases where the fields **E** and **H** are the real parts of complex exponentials such as that given in (1.7). At a given point of space, let us assume that **E** is given by the real part of $\mathbf{E}_0 e^{j\omega t}$, and **H** by the real part of $\mathbf{H}_0 e^{j\omega t}$, where $\mathbf{E}_0$ and $\mathbf{H}_0$ are complex vector functions of position. Let the real part of $\mathbf{E}_0$ be $\mathbf{E}_r$, and the imaginary part $\mathbf{E}_i$, with similar notation for $\mathbf{H}_0$. Then **E** is given by

$$Re(\mathbf{E}_0 e^{j\omega t}) = Re[(\mathbf{E}_r + j\mathbf{E}_i)(\cos \omega t + j \sin \omega t)] = \mathbf{E}_r \cos \omega t - \mathbf{E}_i \sin \omega t$$

H is given by a similar expression. Poynting's vector is then

$$\mathbf{E} \times \mathbf{H} = (\mathbf{E}_r \times \mathbf{H}_r) \cos^2 \omega t + (\mathbf{E}_i \times \mathbf{H}_i) \sin^2 \omega t \\ - [(\mathbf{E}_r \times \mathbf{H}_i) + (\mathbf{E}_i \times \mathbf{H}_r)] \sin \omega t \cos \omega t.$$

We notice that there are two types of terms: the first two, whose time average is different from zero, since $\cos^2 \omega t$ and $\sin^2 \omega t$ average to $\frac{1}{2}$; and the last term, whose time average is zero, since $\sin \omega t \cos \omega t$ averages to zero. Thus the time average Poynting vector is

$$\text{Average } (\mathbf{E} \times \mathbf{H}) = \tfrac{1}{2}(\mathbf{E}_r \times \mathbf{H}_r + \mathbf{E}_i \times \mathbf{H}_i). \tag{5.1}$$

This can be rewritten in a convenient way, by using the notation of complex conjugates, where the complex conjugate of a complex number is the number obtained from the original one by changing the sign of j wherever it appears, and is indicated by a bar over the number. In terms of this notation, let us consider the quantity $(\mathbf{E} \times \overline{\mathbf{H}})$. This is

$$\begin{aligned}(\mathbf{E} \times \overline{\mathbf{H}}) &= (\mathbf{E}_0 e^{j\omega t}) \times (\overline{\mathbf{H}}_0 e^{-j\omega t}) \\ &= (\mathbf{E}_0 \times \overline{\mathbf{H}}_0) = (\mathbf{E}_r \times j\mathbf{E}_i) \times (\mathbf{H}_r - j\mathbf{H}_i) \\ &= (\mathbf{E}_r \times \mathbf{H}_r + \mathbf{E}_i \times \mathbf{H}_i) + j(\mathbf{E}_i \times \mathbf{H}_r - \mathbf{E}_r \times \mathbf{H}_i).\end{aligned} \tag{5.2}$$

We see that, except for the factor $\frac{1}{2}$, the real part of (5.2) is just the same as the quantity appearing in (5.1). That is, we have

$$\text{Average } (\mathbf{E} \times \mathbf{H}) = \tfrac{1}{2} Re(\mathbf{E} \times \overline{\mathbf{H}}), \tag{5.3}$$

where the **E** and **H** appearing on the right side of the equation are the complex quantities whose real parts give the real **E** and **H** appearing on the left of the equation. A similar derivation for the energy density shows that

$$\text{Average } \tfrac{1}{2}\epsilon E^2 = \tfrac{1}{4}\epsilon \mathbf{E} \cdot \overline{\mathbf{E}}, \tag{5.4}$$

with a similar formula for the magnetic energy.

6. Power Flow and Energy Density in a Plane Wave.—In (2.2) we have found that **E** and **H** in a plane wave are at right angles to each other, and at right angles to the direction of propagation. In particular, we have for a nonabsorbing medium, where Z_0 is real,

$$\frac{1}{2} Re(\mathrm{E} \times \overline{\mathrm{H}}) = \frac{1}{2}\frac{\mathrm{k}}{Z_0}(\mathrm{E} \cdot \overline{\mathrm{E}}) = \frac{1}{2}\mathrm{k}Z_0(\mathrm{H} \cdot \overline{\mathrm{H}}). \tag{6.1}$$

That is, the flow of energy is along the direction of propagation, and proportional to the square of the amplitude of **E** or of **H**. Similarly the average electrical energy density is $\frac{1}{4}\epsilon(\mathrm{E} \cdot \overline{\mathrm{E}})$, and the average magnetic density is $\frac{1}{4}\mu(\mathrm{H} \cdot \overline{\mathrm{H}})$. Because of the relation (2.1) between the amplitudes of **E** and **H**, we find easily that the average magnetic energy equals the average electrical energy, so that the total energy density is $\frac{1}{2}\epsilon(\mathrm{E} \cdot \overline{\mathrm{E}})$. We note, from (6.1) and (2.1), together with (1.8), that we have the relation

$$\text{Energy flux} = v \times \text{energy density}. \tag{6.2}$$

This equation has a simple meaning. If the energy were flowing with velocity v, in the direction of propagation of the wave, all the energy contained in a cylinder of unit cross section, and height equal to v, would cross unit cross section per second, forming the flux.

In a conducting medium, in which α is different from zero, **E** and **H** will each contain a factor $e^{-\alpha z}$, so that Poynting's vector and the energy density will each have a factor $e^{-2\alpha z}$, showing that the intensity is damped or attenuated by this factor in traveling along. Poynting's theorem of course tells us that the energy lost to the wave as it advances is used up in resistive heating of the medium. We readily find that, in a conducting medium, the electrical energy is no longer equal to the magnetic energy, but as the conductivity becomes greater and greater, the electric energy becomes smaller and smaller compared with the magnetic energy, finally vanishing in the limit of infinite conductivity. We shall investigate these relations more completely in the next chapter.

Problems

1. If the generation of heat per unit volume in a conductor carrying a current is σE^2, prove that, for a cylindrical conductor of resistance R, carrying a current i, the rate of generation is i^2R.

2. Given a cylindrical wire carrying a current. Find the values of E and H on the surface of the wire, computing Poynting's vector, and show that it represents a flow of energy into the wire. Show that the amount flowing into a given length of wire is just enough to supply the energy that appears as heat in the length. Note that the surface of a wire carrying current is not an equipotential, so that there can be a component of electric field parallel to it.

3. Calculate the electrical and magnetic energies in a plane wave traveling in a conductor, and show by direct comparison that they are different from each other. What happens in the limiting cases $\sigma \to 0$ and $\sigma \to \infty$, that is, insulators and perfect conductors?

4. In the absence of charges, solutions of Maxwell's equations may be obtained from a vector potential **A** alone, setting the scalar potential equal to zero. Work this out for the general case of plane waves propagated along the z axis, showing that the vector potential can have two arbitrary complex amplitudes for its x and y components, the z dependence being in the factor $e^{j\omega t - \gamma z}$. From this find the relations between **E** and **H** in a plane wave, showing that your results are identical with those obtained in the text.

5. A coaxial line of length l, inner and outer radii a and b, carries a steady current I. The resistance of the inner conductor is R_i, that of the outer conductor R_0, and the load resistance is R. Find expressions for the components of **E** and **H** at any point between the conductors. Show that the component of E parallel to the axis reverses direction as one moves from the inner to the outer conductor, and that it is zero at a radius r_0 given by $\ln (r_0/a) = [R_i/(R_0 + R_i)] \ln (b/a)$. Compute the Poynting vector, and discuss the power flow in this field. Integrate the radial component of the Poynting vector over the surface of both inner and outer conductor, and show that these integrals yield the I^2R losses in each. Apply the Poynting theorem to a volume bounded by the planes z_1 and z_2 and the cylindrical surfaces r_1 and r_2.

6. Consider a plane wave in empty space given by

$$E_x = \sqrt{(\mu_0/\epsilon_0)}\, H_y = E_0 e^{j(\omega t - \beta z)},$$

with $\beta = \omega \sqrt{\epsilon_0 \mu_0}$. Show that an approximate solution to Maxwell's equations can be obtained to represent a plane wave of finite cross section by considering E_0 to be a slowly varying function of x and y and adding to the above field components the longitudinal components

$$E_z = -\frac{j}{\beta}\frac{\partial E_0}{\partial x} e^{j(\omega t - \beta z)} \qquad \text{and} \qquad H_z = -\frac{j}{\beta}\sqrt{\frac{\epsilon_0}{\mu_0}}\frac{\partial E_0}{\partial y} e^{j(\omega t - \beta z)}.$$

By slowly varying is meant that the percentage change in the fields is small compared with unity if one moves a distance of one wave length in a direction transverse to the direction of propagation, so that second derivatives and products of first derivatives with respect to x and y may be neglected.

7. Prove that, if a piece of steel is magnetized by an alternating current, the energy loss per cycle due to hysteresis is given by $\oint \mathbf{H} \cdot d\mathbf{B}$ per unit volume.

8. A magnetic circuit of uniform cross section consists of two sections separated by air gaps of length so small that fringing may be neglected. Consider the process of increasing the air-gap lengths each by an infinitesimal amount, maintaining the magnetizing current constant. Compute the change in magnetic field energy and the work done by the emf that maintains the magnetizing current constant during the displacement. From this obtain the mechanical work done in effecting the displacement, and show that the force with which the two sections attract each other is given by B^2A/μ_0, where A is the cross section of the magnetic circuit.

CHAPTER IX

ELECTRON THEORY AND DISPERSION

In Chap. VIII we investigated the propagation of plane waves in a medium with a given dielectric constant and conductivity. We have seen that in a nonconducting nonmagnetic medium a wave is propagated with a velocity c/n, where c is the velocity of light in empty space, n the index of refraction, which is equal to $\sqrt{\kappa_e}$. In a conducting medium, there is a damped wave, whose propagation constant, determining both the velocity of propagation and the rate of absorption, is given by Eq. (1.5), Chap. VIII, predicting a definite variation of propagation properties with frequency. These two formulas give a very straightforward way of testing the electromagnetic theory of light: we can measure the dielectric constant and conductivity of a material, and see if its optical index of refraction and absorption coefficient are properly related to these constants. One of the first observations after the formulation of Maxwell's theory was that these relations are not fulfilled by real substances. The departures between observation and the simple theory have been very useful in gaining a knowledge of the structure of actual dielectrics and conductors. Briefly, the discrepancy between theory and experiment is explained by supposing that the dielectric constant and conductivity are functions of frequency; a large part of the electron theory of solids is devoted to an explanation of the nature of this frequency variation.

The experimental situation is much clearer than it was in Maxwell's day, because of the wider ranges of the spectrum that have been explored in the meantime. We shall first describe the situation for dielectrics without conductivity. For long radio wave lengths, waves can be propagated with a velocity given by the index of refraction determined from the static dielectric constant. As the frequency increases, however, the index of refraction starts to increase, until somewhere in the spectrum it goes through a maximum, then drops suddenly, and begins to increase again. The phenomenon of change of index of refraction with frequency is called "dispersion," and the sudden drop we have just mentioned is called "anomalous dispersion." In the neighborhood of a region of anomalous dispersion, there is absorption, even when the material does not absorb elsewhere; the

3. Calculate the electrical and magnetic energies in a plane wave traveling in a conductor, and show by direct comparison that they are different from each other. What happens in the limiting cases $\sigma \to 0$ and $\sigma \to \infty$, that is, insulators and perfect conductors?

4. In the absence of charges, solutions of Maxwell's equations may be obtained from a vector potential **A** alone, setting the scalar potential equal to zero. Work this out for the general case of plane waves propagated along the z axis, showing that the vector potential can have two arbitrary complex amplitudes for its x and y components, the z dependence being in the factor $e^{j\omega t - \gamma z}$. From this find the relations between **E** and **H** in a plane wave, showing that your results are identical with those obtained in the text.

5. A coaxial line of length l, inner and outer radii a and b, carries a steady current I. The resistance of the inner conductor is R_i, that of the outer conductor R_0, and the load resistance is R. Find expressions for the components of **E** and **H** at any point between the conductors. Show that the component of E parallel to the axis reverses direction as one moves from the inner to the outer conductor, and that it is zero at a radius r_0 given by $\ln(r_0/a) = [R_i/(R_0 + R_i)] \ln(b/a)$. Compute the Poynting vector, and discuss the power flow in this field. Integrate the radial component of the Poynting vector over the surface of both inner and outer conductor, and show that these integrals yield the I^2R losses in each. Apply the Poynting theorem to a volume bounded by the planes z_1 and z_2 and the cylindrical surfaces r_1 and r_2.

6. Consider a plane wave in empty space given by

$$E_x = \sqrt{(\mu_0/\epsilon_0)}\, H_y = E_0 e^{j(\omega t - \beta z)},$$

with $\beta = \omega \sqrt{\epsilon_0 \mu_0}$. Show that an approximate solution to Maxwell's equations can be obtained to represent a plane wave of finite cross section by considering E_0 to be a slowly varying function of x and y and adding to the above field components the longitudinal components

$$E_z = -\frac{j}{\beta}\frac{\partial E_0}{\partial x} e^{j(\omega t - \beta z)} \qquad \text{and} \qquad H_z = -\frac{j}{\beta}\sqrt{\frac{\epsilon_0}{\mu_0}}\frac{\partial E_0}{\partial y} e^{j(\omega t - \beta z)}.$$

By slowly varying is meant that the percentage change in the fields is small compared with unity if one moves a distance of one wave length in a direction transverse to the direction of propagation, so that second derivatives and products of first derivatives with respect to x and y may be neglected.

7. Prove that, if a piece of steel is magnetized by an alternating current, the energy loss per cycle due to hysteresis is given by $\oint \mathbf{H} \cdot d\mathbf{B}$ per unit volume.

8. A magnetic circuit of uniform cross section consists of two sections separated by air gaps of length so small that fringing may be neglected. Consider the process of increasing the air-gap lengths each by an infinitesimal amount, maintaining the magnetizing current constant. Compute the change in magnetic field energy and the work done by the emf that maintains the magnetizing current constant during the displacement. From this obtain the mechanical work done in effecting the displacement, and show that the force with which the two sections attract each other is given by B^2A/μ_0, where A is the cross section of the magnetic circuit.

CHAPTER IX

ELECTRON THEORY AND DISPERSION

In Chap. VIII we investigated the propagation of plane waves in a medium with a given dielectric constant and conductivity. We have seen that in a nonconducting nonmagnetic medium a wave is propagated with a velocity c/n, where c is the velocity of light in empty space, n the index of refraction, which is equal to $\sqrt{\kappa_e}$. In a conducting medium, there is a damped wave, whose propagation constant, determining both the velocity of propagation and the rate of absorption, is given by Eq. (1.5), Chap. VIII, predicting a definite variation of propagation properties with frequency. These two formulas give a very straightforward way of testing the electromagnetic theory of light: we can measure the dielectric constant and conductivity of a material, and see if its optical index of refraction and absorption coefficient are properly related to these constants. One of the first observations after the formulation of Maxwell's theory was that these relations are not fulfilled by real substances. The departures between observation and the simple theory have been very useful in gaining a knowledge of the structure of actual dielectrics and conductors. Briefly, the discrepancy between theory and experiment is explained by supposing that the dielectric constant and conductivity are functions of frequency; a large part of the electron theory of solids is devoted to an explanation of the nature of this frequency variation.

The experimental situation is much clearer than it was in Maxwell's day, because of the wider ranges of the spectrum that have been explored in the meantime. We shall first describe the situation for dielectrics without conductivity. For long radio wave lengths, waves can be propagated with a velocity given by the index of refraction determined from the static dielectric constant. As the frequency increases, however, the index of refraction starts to increase, until somewhere in the spectrum it goes through a maximum, then drops suddenly, and begins to increase again. The phenomenon of change of index of refraction with frequency is called "dispersion," and the sudden drop we have just mentioned is called "anomalous dispersion." In the neighborhood of a region of anomalous dispersion, there is absorption, even when the material does not absorb elsewhere; the

index of refraction and absorption coefficient in this neighborhood behave much like the reactance and resistance of a resonant electric circuit. Most materials have several regions of anomalous dispersion and absorption, in different parts of the spectrum. Many materials, for instance, have absorption in the infrared, and practically all materials absorb in the ultraviolet. Materials that absorb in the visible part of the spectrum are those which appear colored. By the time we go through the ultraviolet to the X-ray region, anomalous dispersion has stopped, and the index of refraction of all materials has become almost exactly unity; it is for this reason that lenses and prisms are practically impossible in the X-ray region. The absorption has also largely decreased in the X-ray region, and it is for that reason that X rays can penetrate so many types of materials. Passing beyond, to the region of gamma rays, all materials become rather transparent and nonrefracting.

For conductors, we found in Eq. (1.5), Chap. VIII, that γ is given by $j\omega\sqrt{(\epsilon - j\sigma/\omega)\mu}$. We see that as the frequency approaches zero the second term in the radical becomes infinitely greater in magnitude than the first. Thus at low frequencies the optical properties of a conductor are determined entirely by the conductivity, and not by the dielectric constant; in fact, there is no experimental way of finding the dielectric constant of a good conductor at low frequencies. The values of σ for good conductors, such as metals, are such that σ/ϵ does not become comparable with ω until we reach frequencies in the visible part of the spectrum. It is found experimentally that the optical properties of metals are well described by this simple theory through the radio-frequency, microwave, and infrared parts of the spectrum, and that, as we should expect, the dielectric effects begin to be important in the visible region. The situation there, however, is much more complicated than would be indicated by Eq. (1.5), Chap. VIII; for by then, there is good evidence that both ϵ and σ vary in complicated ways with the frequency. The variation of ϵ for a metal is not unlike that for a dielectric; there is anomalous dispersion, and consequent absorption, superposed on the conductivity. As we go through the ultraviolet to the X-ray region, the effect of conductivity becomes negligible, and metals are no more absorbing than other materials.

We shall now examine the way in which simple electronic models of dielectrics and conductors can lead to variations of dielectric constant and conductivity with frequency which are in at least qualitative agreement with the observations.

1. Dispersion in Gases.—In Chap. IV, we described the nature of dielectric polarization, stating that it can arise in two ways: from the polarization of the atoms or molecules, by the displacement of the electrons in them; and by orientation of dipoles already existing, under the action of the external field. We shall discuss only the first type of polarization in the present chapter. We shall find in a later section that, in a solid or liquid, the problem of polarization is complicated by the interactions between the molecules or atoms; but in a gas, the molecules are far enough apart so that we can neglect the interactions between them. Each molecule contains charges that can be displaced under the action of an external field, and these charges act as if they were held to positions of equilibrium by restoring forces proportional to the displacement. Thus in a static case an electron of charge e is acted on by the force eE of the external electric field, and $-ax$, a linear restoring force, proportional to the displacement x, with a constant of proportionality a. The displacement is then $x = (e/a)E$, and the induced dipole moment $ex = (e^2/a)E$. Thus the polarizability α of a molecule is e^2/a, and using Eq. (2.4) of Chap. IV, the dielectric constant is given by

$$\kappa_e = 1 + N\frac{e^2}{a\epsilon_0}, \tag{1.1}$$

where N is the number per unit volume.

The value (1.1) is the static value. If the external field varies with time, we must take account of the fact that the electron has a mass m, thus possessing inertia. We shall also introduce, in addition to the external electric force and the elastic restoring force, a damping force proportional to the velocity, to account for the absorption. The equation of motion for the electron is then

$$m\frac{d^2x}{dt^2} + mg\frac{dx}{dt} + m\omega_0^2 x = eE, \tag{1.2}$$

where we have rewritten the linear restoring force in terms of a constant $m\omega_0^2$. If E varies sinusoidally with time, as the real part of an expression containing the exponential factor $e^{j\omega t}$, we can then solve (1.2), as always in determining the forced oscillations of a linear oscillator, by assuming that x varies also as $e^{j\omega t}$. We then find

$$x = \frac{(e/m)E}{\omega_0^2 - \omega^2 + j\omega g}.$$

That is, the electron vibrates with the same frequency as the external field, but with an amplitude depending on the frequency. If we have

N_k electrons per unit volume characterized by constants ω_k and g_k, the polarization is

$$P = E \sum_k \frac{N_k e^2/m}{\omega_k^2 - \omega^2 + j\omega g_k},$$

and the dielectric constant is

$$\kappa_e = 1 + \sum_k \frac{N_k e^2/m\epsilon_0}{\omega_k^2 - \omega^2 + j\omega g_k}. \tag{1.3}$$

In the limit of low frequencies, where ω can be neglected, this reduces essentially to the value (1.1), the static dielectric constant, but at each of the resonant frequencies ω_k there is a phenomenon such as is found with the impedance of an electric circuit, near its resonant frequency. The magnitude of the dielectric constant varies rapidly with frequency, and at the same time it becomes complex, leading to absorption, as well as dispersion.

It is convenient, in an absorbing medium, to introduce an absorption coefficient k, as well as an index of refraction n, by the relation

$$\gamma = j\frac{\omega}{c}(n - jk), \qquad e^{-\gamma z} = e^{-(\omega/c)kz}e^{-j(\omega/c)nz}.$$

Then we have

$$(n - jk)^2 = \kappa_e.$$

For a gas, there are few enough molecules so that the second term of (1.3) is small compared with unity. Thus we have approximately

$$n - jk = 1 + \frac{1}{2}\sum_k \frac{N_k e^2/m\epsilon_0}{\omega_k^2 - \omega^2 + j\omega g_k},$$

and, if we separate into real and imaginary parts, we obtain

$$n = 1 + \frac{1}{2}\sum_k \frac{(N_k e^2/m\epsilon_0)(\omega_k^2 - \omega^2)}{(\omega_k^2 - \omega^2)^2 + \omega^2 g_k^2}$$

$$k = \frac{1}{2}\sum_k \frac{(N_k e^2/m\epsilon_0)\omega g_k}{(\omega_k^2 - \omega^2)^2 + \omega^2 g_k^2}.$$

If we consider these two quantities as functions of frequency, they show the properties we have already discussed. As the frequency increases, the index of refraction goes through a phenomenon of anomalous dispersion in the neighborhood of each of the resonance

frequencies ω_k, the index eventually approaching unity, and being in fact slightly less than unity, for very large values of frequency; while the absorption coefficient is small everywhere except near each of the resonance frequencies. The behavior of n and k near resonance is as shown in Fig. 24.

2. Dispersion in Liquids and Solids.—In the case of solids and liquids we may no longer make the approximation that the force acting on an electron is simply the electric vector of the electromagnetic wave in free space, but must take into account the added force on the electron due to the polarization of the body. We can calculate this force as follows: we imagine a small sphere of radius R (with its center at the position of the electron in question) cut out of the medium. If we do this, without disturbing the distribution of polarization outside the sphere, we have induced charges on the surface of this spherical volume from which we calculate the force at the center of the sphere. The surface density of induced charge on a spherical ring at an angle θ to the direction of the field is $P \cos \theta$. Since the area of a ring included between angles θ and $\theta + d\theta$ is $2\pi R^2 \sin \theta \, d\theta$, the charge on the ring is $2\pi P R^2 \cos \theta \sin \theta \, d\theta$. This charge produces a field at the center of the sphere whose component parallel to E is

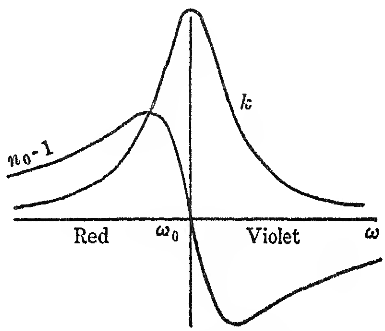


FIG. 24.—Anomalous dispersion, showing index of refraction and absorption coefficient as function of frequency.

$$dE_1 = \frac{2\pi P R^2 \cos^2 \theta \sin \theta \, d\theta}{4\pi\epsilon_0 R^2},$$

so that the total charge on the spherical surface produces a field at the center equal to

$$E_1 = \frac{P}{2\epsilon_0} \int_0^\pi \cos^2 \theta \sin \theta \, d\theta = \frac{P}{3\epsilon_0}.$$

The total electric field at the center of this sphere is then

$$E + \frac{P}{3\epsilon_0}.$$

Of course, there is still the contribution to the force by the atoms inside the little sphere we have cut out, but in an isotropic medium this averages to zero.

We can now carry over our calculations for gases if we replace E by $(E + P/3\epsilon_0)$ in the expression for x. Thus we get

$$P = \left(E + \frac{P}{3\epsilon_0}\right) \sum_k \frac{N_k e^2/m}{\omega_k^2 - \omega^2 + j\omega g_k},$$

and using the relations $D = \kappa_e \epsilon_0 E = \epsilon_0 E + P$, we have

$$E + \frac{P}{3\epsilon_0} = \frac{\kappa_e + 2}{3} E,$$

and we find for κ_e

$$\frac{\kappa_e - 1}{\kappa_e + 2} = \frac{(n - jk)^2 - 1}{(n - jk)^2 + 2} = \frac{1}{3} \sum_k \frac{N_k e^2/m\epsilon_0}{\omega_k^2 - \omega^2 + j\omega g_k}.$$

If N represents the number of atoms per unit volume, and if f_k is the fraction of atoms of the kth type, so that $N_k = f_k N$, we may rewrite it

$$\frac{(n - jk)^2 - 1}{(n - jk)^2 + 2} \frac{1}{N} = \frac{1}{3} \sum_k \frac{f_k e^2/m\epsilon_0}{\omega_k^2 - \omega^2 + j\omega g_k}. \tag{2.1}$$

In this formula, the quantities f_k, which are independent of the number of atoms per unit volume, are often called the "oscillator strengths" of the various resonances.

For a transparent substance, where we can neglect the damping force, and the index of refraction is real, we have for a given frequency of light

$$\frac{n^2 - 1}{n^2 + 2} \frac{1}{\rho_0} = \text{constant},$$

where ρ_0 is the density of the body, obviously proportional to N. This law, known as the Lorenz-Lorentz law, is surprisingly well obeyed for many substances. In many cases it even gives approximately correctly the relationship between the index of refraction of a liquid and of its vapor. In the limit of very long electromagnetic waves, and for the electrostatic case, we have

$$\frac{\kappa_e - 1}{\kappa_e + 2} \frac{1}{\rho_0} = \text{constant},$$

the so-called "Clausius-Mosotti relation" between dielectric constant and density.

There is a different way of handling dispersion in liquids and solids, equivalent to this, but exhibiting the result in different form.

We shall consider only the case where there is but one type of oscillating electron. In (1.2) we replace the right side by $e(E + P/3\epsilon_0)$, but also we write $P = Nex$, which would be the relation for only one type of electron, and incorporate the term in x in the left side of the equation, so that the equation becomes

$$m\frac{d^2x}{dt^2} + mg\frac{dx}{dt} + m\left(\omega_0^2 - \frac{Ne^2}{3\epsilon_0 m}\right)x = eE.$$

Solving as in (1.3), we then find

$$(n - jk)^2 = 1 + \frac{Ne^2/m_0}{\bar{\omega}_0^2 - \omega^2 + j\omega g}, \tag{2.2}$$

where

$$\bar{\omega}_0^2 = \omega_0^2 - \frac{Ne^2}{3\epsilon_0 m}. \tag{2.3}$$

That is, we have the same type of anomalous dispersion that we found in (1.3) for a gas, but with the resonant frequency displaced according to (2.3). If there are various resonant frequencies, it is no longer a simple matter to prove that a solution like (1.3) can be set up, but the proof can be given, by methods similar to those used in discussing the normal coordinates of a coupled vibrating system in classical mechanics. We note that, since N is proportional to the density, the resonant frequencies appearing in a formula of the type of (2.2) vary with the density, whereas the frequencies appearing in the Lorenz-Lorentz formula (2.1) are independent of density.

3. Dispersion in Metals.—The simple electron theory of metals assumes that free electrons wander about among fixed ions, and carry the current. On the average there is no resultant force acting on the electrons, except the external field, so that there is a force eE acting on them. The equation of motion is thus similar to (1.2), except that there is no restoring force, so that ω_0 is zero. We must assume a damping force, proportional to the velocity, and we find that it is simply related to the conductivity, according to Ohm's law. For in the case of a constant external field, the electrons will acquire constant velocity, which by (1.2) will be eE/mg, so that, if N is the number of conducting electrons per unit volume, the current density is

$$J = Nev = \frac{Ne^2E}{mg} = \sigma E, \qquad \sigma = \frac{Ne^2}{mg}. \tag{3.1}$$

This holds only for a steady field, however. If E varies as $e^{j\omega t}$, then we find from (1.2) that

$$\sigma = \frac{Ne^2}{mg + mj\omega},$$

so that the conductivity varies with frequency, vanishing with infinite frequency. In considering propagation, we have seen in Eq. (1.5), Chap. VIII, that the combination $\kappa_e - j\sigma/\omega\epsilon_0$ is to be used in place of the dielectric constant in a transparent medium. Using (1.3) for κ_e, this gives

$$\kappa_e - \frac{j\sigma}{\omega\epsilon_0} = (n - jk)^2 = 1 + \frac{Ne^2/m\epsilon_0}{-\omega^2 + j\omega g} + \sum_k \frac{N_k e^2/m\epsilon_0}{\omega_k^2 - \omega^2 + j\omega g_k}, \quad (3.2)$$

in which it is clear that the term coming from the conductivity has just the same form as the other terms, except for having its resonant frequency equal to zero. Separating real and imaginary parts of (3.2), and using (3.1), where we shall now use σ for its value at zero frequency, we may rewrite (3.2) in the form

$$n^2 - k^2 = 1 - \frac{\sigma}{g\epsilon_0}\frac{1}{1 + \omega^2/g^2} + \sum_k \frac{(N_k e^2/m\epsilon_0)(\omega_k^2 - \omega^2)}{(\omega_k^2 - \omega^2) + \omega^2 g_k^2}$$

$$2nk = \frac{\sigma}{\omega\epsilon_0}\frac{1}{1 + \omega^2/g^2} + \sum_k \frac{(N_k e^2/m\epsilon_0)(\omega g_k)}{(\omega_k^2 - \omega^2)^2 + \omega^2 g_k^2}.$$

We notice that as the frequency becomes low compared with σ/ϵ_0, which for good conductors is in the ultraviolet part of the spectrum, the first term in the product nk becomes large compared with unity, masking the effect of the bound electrons. The difference $n^2 - k^2$ does not become correspondingly large, so that in the limit of low frequencies, n becomes equal to k, and both approach $\sqrt{\sigma/2\omega\epsilon_0}$, neglecting ω compared with g. However, it is only at low frequencies that these simplifications occur. As the frequency enters the near infrared or visible region, it becomes of the same order of magnitude as both σ/ϵ_0 and g, so that the contributions of the free electrons become complicated, and at the same time nk decreases so that the contributions of the bound electrons become important. It is thus natural that experimentally the curves of $n^2 - k^2$ and nk throughout the visible part of the spectrum are very complicated, though they can be fitted fairly accurately with formulas of the type we have derived, assuming bound as well as free electrons. In the ultraviolet, the contributions of the free electrons become small compared with those of the bound electrons having resonance in that region, and a metal does not behave essentially differently from an insulator.

4. The Quantum Theory and Dispersion.—The picture of dispersion and of the optical properties of dielectrics and metals that

we have presented is based on simple classical models of the behavior of electrons in these materials. We have assumed that dielectrics contain electrons held to positions of equilibrium by linear restoring forces, subject to resistances proportional to their velocity, and we have assumed the conduction electrons in a metal to be similar, except that they have no restoring forces, but only resistance. The theory in this form was developed by Drude and Lorentz, in the early days of the electron theory. It has proved in practice to be so good as to surprise one, when the crudity of the assumptions is considered. The real electrons in atoms, molecules, and metals, as we now know from the quantum theory, behave very differently from our simple picture, but it is a remarkable fact that the quantum theory leads essentially to a justification of our mathematical formulation, though not of the simple hypotheses underlying it. According to the quantum theory, as is well known, atoms, molecules, and other systems have certain energy levels in which they can exist, and the emission and absorption of radiation are associated with transitions between energy levels, the frequency ν associated with two levels E_1 and E_2 being given by Planck's relation $E_2 - E_1 = h\nu$, where h is Planck's constant. Atoms and molecules have only a discrete set of excited energy levels, so that from the ground state they can absorb a discrete set of definite frequencies. Metals, on the other hand, have a continuum of excited levels, associated with the free electrons.

In the quantum theory, we can investigate the behavior of an atom, molecule, or metal, under the action of an external radiation field. We find, by application of certain perturbation methods, that the effect of the atom on the field can be replaced by an equivalent set of linear oscillators, associated with the various transitions that are allowed by Planck's relation. Each oscillator has a resonant frequency given by one of the allowed transitions. With N atoms per unit volume, however, we do not have the equivalent of N oscillating electrons per unit volume of each of the frequencies; we have instead only Nf, where f is a fraction, sometimes called the "oscillator strength," and which can be calculated by quantum mechanics. We further find that the oscillators have not only resonant frequencies, but also damping terms, which can be correlated with the broadening of the upper and lower states, by collisions with other atoms and other perturbing processes. Thus the net result of the quantum discussion is a theory mathematically equivalent to the one developed in the present chapter, but with quite a different physical interpretation. It is for this reason that a purely classical discussion of dis-

persion, such as we have given, is of real physical importance, and not merely an academic matter.

In metals, the quantum theory also leads to quite a different picture from an elementary classical theory. Electrons are not governed by classical statistics, but by Fermi statistics, as a result of which the electrons in a metal are not at rest in the absence of a field, as our classical equation of motion would suggest, but actually are in continuous motion with a very high velocity. In the absence of an external field, however, as many are traveling in one direction as in the opposite direction, so that there is no net current. In the presence of a field, the electrons gain an average acceleration that is essentially what corresponding classical electrons would experience, though in most cases they have an effective mass that is different from their classical mass, because of the structure of the energy bands of the metal. The electrons, once they are accelerated, are subject, in quantum theory as in our classical picture, to a frictional resistive force, but we can give a physical explanation of the friction, instead of merely postulating it. The electrons have many of the properties of waves, as is observed experimentally in electron diffraction, and these waves are scattered by the irregularities in the metal produced by the thermal oscillations of the atoms. It is this scattering which produces the effect of friction, dissipating any average momentum that the electrons may acquire. Thus the quantum theory eventually arrives at a picture of the electrons in a metal that is not unlike that of elementary classical theory, and the quantum theory of the interaction of metals with a radiation field essentially verifies the simple theory that we have described in this chapter.

This brief discussion of the relation of the classical theory of dispersion to the quantum theory is not expected to give a clear idea of the quantum phenomena to one not already familiar with them, but is intended merely to make it plain that, as we have already mentioned, the classical theory of dispersion is of great importance physically, furnishing a description of experimental facts that in its broad outlines is correct, and correlates with the quantum theory.

Problems

1. Show that, in the case of normal dispersion for the visible spectrum where there is an absorption band in the ultraviolet, the index of refraction can be written as

$$n^2 = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots,$$

where λ is the wave length in vacuum and A, B, C are constants.

If there is also absorption in the infrared, show that the index of refraction is then given by

$$n^2 = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots - A'\lambda^2 - B'\lambda^4 \cdots .$$

2. Measurements of H_2 gas give the following values of the index of refraction:

λ in A	$(n - 1)10^7$
5,462.260	1,396
4,078.991	1,426
3,342.438	1,461
2,894.452	1,499
2,535.560	1,547
2,302.870	1,594
1,935.846	1,718
1,854.637	1,760

Using the expression in Prob. 1 for n^2 in reciprocal powers of λ, calculate the best values of A, B, and C. If the measurements are made at room temperature and atmospheric pressure, calculate the resonant frequency ω_0 and wave length from these constants.

3. Prove that in the case of anomalous dispersion for gases the maximum and minimum values of n occur at the positions where the absorption coefficient reaches half its maximum value. Find the relation between g_k and the half width of the absorption band. Assume $g_k/\omega_k \ll 1$.

4. For the D line of sodium the following values of the constants in the dispersion formula are found:

$$\omega_0 = 3 \times 10^{15}; \qquad g = 2 \times 10^{10}; \qquad \frac{Ne^2}{m\epsilon_0} = 10^{23}.$$

Plot the index of refraction n and the absorption coefficient k as a function of the frequency of the light. Find the maximum and minimum values of the index of refraction n. Find the maximum value of the absorption coefficient k and the half width of the absorption band in angstrom units.

5. Show that for gases the Lorenz-Lorentz law takes the approximate form $\frac{2}{3}\frac{n-1}{\rho_0}$ = constant. The following measurements have been made on air (ρ_0 given in arbitrary units):

ρ_0	n
1.00	1.0002929
14.84	1.004338
42.13	1.01241
69.24	1.02044
96.16	1.02842
123.04	1.03633
149.53	1.04421
176.27	1.05213

Calculate $\frac{2}{3}\frac{n-1}{\rho_0}$ and $\frac{n^2-1}{n^2+2}\frac{1}{\rho_0}$ for each of these measurements, and compare the constancy of the results (calculate to four significant figures).

6. The indices of refraction for the sodium D line, and densities in grams per cubic centimeter of some liquids at 15°C are

Liquid	ρ_0	n
Water	0.9991	1.3337
Carbon bisulphide	1.2709	1.6320
Ethyl ether	0.7200	1.3558

Calculate the indices of refraction for the vapors at 0°C and 760 mm pressure. The observed values for the vapors are 1.000250, 1.00148, and 1.00152, respectively.

7. The quantity $\dfrac{m(n^2-1)}{(n^2+2)\rho_0}$ is called the "refractivity" of a substance if m denotes its mass. Prove that the refractivities of mixtures of substances equal the sum of the refractivities of the constituents. (Neglect damping forces from the start.)

8. Show that the molecular refractivity of a compound, defined as $\dfrac{n^2-1}{n^2+2}\dfrac{M}{\rho_0}$, where M is the molecular weight, is equal to the sum of the atomic refractivities of the atoms of which the compound is formed. (Neglect damping forces.)

9. For the following gases we have the following values of $(n-1)_\infty$ extrapolated to long wave lengths:

Gas	$(n-1)_\infty \cdot 10^6$
H_2	136.35
N_2	294.5
O_2	265.3

Calculate the values of $(n-1)_\infty$ for the following gases: H_2O, NH_3, NO, N_2O_4, O_3. The measured values are 245.6, 364.6, 288.2, 496.5, 483.6, all times 10^6. Find the percentage discrepancy between the calculated and observed values.

CHAPTER X

REFLECTION AND REFRACTION OF ELECTROMAGNETIC WAVES

In Chap. VIII we investigated the behavior of a plane wave in a homogeneous medium, and showed that it is propagated with a velocity equal to the velocity in free space, divided by the index of refraction. Next in Chap. IX we investigated the physical nature of various types of media, and found the features leading to various indices of refraction, their variation with wave length, and to the absorption that often accompanies dispersion. Now we shall consider what happens at an infinite plane boundary between two semi-infinite media of different indices of refraction, such for instance as free space and a dielectric, or free space and a metallic conductor. We shall be led to the familiar laws of reflection and refraction, laws that were established, before the electromagnetic theory, merely from general wave theory. To consider the behavior at the boundary, we must find the boundary conditions holding at a surface of discontinuity between two media.

1. Boundary Conditions at a Surface of Discontinuity.—In each of our two media, we assume that the solution of Maxwell's equations that we desire is a plane wave, just as in an infinite medium. At the boundary between the media, however, certain boundary conditions are to be met, and these demand that there be definite relations between the waves in the two media. These conditions have been derived earlier, in Eq. (3.3) of Chap. IV for the electric vectors, and in Eq. (3.2) of Chap. VI for the magnetic ones. We rewrite them:

$$\begin{aligned}&\text{Normal components of } \mathbf{D} \text{ and } \mathbf{B} \text{ are continuous,}\\ &\text{Tangential components of } \mathbf{E} \text{ and } \mathbf{H} \text{ are continuous,}\end{aligned} \qquad (1.1)$$

at a surface that contains no charge and current. These conditions will prove to be enough to derive the complete laws of reflection and refraction; in fact, the conditions on **D** and **B** follow from the others, and are not necessary for the derivation. We shall find that in general we cannot satisfy them without postulating three separate plane waves: an incident and a reflected wave in one medium, a refracted wave in the other medium. By arguments like those of elementary

optics, involving just the matching up of wave fronts on both sides of the surface of separation, we can prove the simple laws of reflection and refraction: that the angles of incidence and reflection are equal, and Snell's law of refraction. We can go further than this, however, and compute the intensities of the reflected and refracted waves, and hence the reflection coefficients, embodied in Fresnel's equations.

We shall first prove the laws of reflection and refraction, and then Fresnel's equations. Then we shall take up two more complicated cases, which cannot be handled by the most elementary methods: total internal reflection, in which there is an exponentially damped wave of an interesting type in the rarer medium, which does not lead to a flow of power into the medium; and reflection and refraction by a conducting medium such as a metal, in which Snell's law of refraction does not hold in its simple form. In all this discussion, we notice that our treatment holds for any part of the spectrum with equal validity, from the longest electromagnetic or radio waves, through the visible region, to the X rays and gamma rays. The only difference is that the index of refraction and absorption coefficient vary with wave length, in the manner we took up in the preceding chapter. These quantities will appear as constants in our present discussion, however, which will deal throughout with monochromatic waves, of a definite wave length.

2. The Laws of Reflection and Refraction.—Let us assume that the surface of separation is $z = 0$, and that the medium with negative z has constants μ, ϵ, and that for positive values has μ', ϵ', with corresponding indices of refraction and absorption coefficients n, k, and n', k'. Let the incident wave be in the medium with negative z, and let its wave normal have direction cosines l, m, n. Then, in a nonabsorbing medium, which alone we consider in the present section, the exponential factor representing the wave propagation is

$$e^{j\omega[t-(lx+my+nz)/v)]}, \tag{2.1}$$

where $v = c/n$ is the velocity of propagation. We shall simplify by assuming that the wave normal is in the xz plane, and that the angle of incidence, or angle between the wave normal and the z axis, is i; then we have $l = \sin i$, $m = 0$, $n = \cos i$, so that (2.1) may be rewritten

$$e^{j\omega[t-(x \sin i+z \cos i)/v]}. \tag{2.2}$$

Similarly in the medium with positive z, there will be a wave propagated at an angle r with the normal, r being the angle of refraction; its corresponding exponential factor is

$$e^{j\omega[t-(x \sin r+z \cos r)/v']}. \tag{2.3}$$

Along the surface of separation, $z = 0$, the two exponentials must agree; for otherwise, even if we were able to satisfy boundary conditions at one point, these would not hold at other points of the surface. Thus we must have

$$\frac{\sin i}{\sin r} = \frac{v}{v'} = \frac{n'}{n}, \tag{2.4}$$

which is Snell's law of refraction. Our derivation is simply the analytical statement of the elementary fact that the wave fronts in the two media must match each other. In addition to the refracted wave, we can set up in the first medium a reflected wave, whose z component of propagation is reversed, with the exponential

$$e^{j\omega[t-(x \sin i - z \cos i)/v]}. \tag{2.5}$$

Along the plane $z = 0$, this exponential agrees with the other two, so that the boundary conditions can be satisfied; and a little consideration shows that it satisfies the law of reflection, that the angle of reflection equals the angle of incidence i.

3. Reflection Coefficient at Normal Incidence.—We shall next set up the values of the electric and magnetic vectors on both sides of the boundary, use the boundary conditions (1.1), and derive the relations between the amplitudes of incident, reflected, and refracted waves. As a preliminary simple case, let us consider the case of normal incidence, when i and r are zero, so that the exponentials (2.2), (2.3), and (2.5) are independent of x. There will be no normal components of **D** and **B**, since the fields are transverse to the direction of propagation, which in this case is the z axis. In the incident wave, let **E** be along the x axis, **H** along the y axis. From Eq. (2.1) of Chap. VIII, we have

$$\frac{E_x}{H_y} = Z_0 = \sqrt{\frac{\mu}{\epsilon}}. \tag{3.1}$$

Similarly for the refracted wave we assume **E** is along x, **H** along y, and have

$$\frac{E'_x}{H'_y} = Z'_0 = \sqrt{\frac{\mu'}{\epsilon'}}. \tag{3.2}$$

For the reflected wave, either **E** or **H** must be reversed in phase with respect to the incident wave, since the wave travels in the opposite direction. Thus we have

$$\frac{E''_x}{H''_y} = -\sqrt{\frac{\mu}{\epsilon}}. \tag{3.3}$$

Our boundary conditions may then be stated with respect to **E**:

$$E_x + E''_x = E'_x, \qquad \sqrt{\frac{\epsilon}{\mu}}\,(E_x - E''_x) = \sqrt{\frac{\epsilon'}{\mu'}}\,E'_x.$$

Eliminating E'_x, we find for the ratio of reflected to incident amplitude

$$\frac{E''_x}{E_x} = \frac{\sqrt{\epsilon/\mu} - \sqrt{\epsilon'/\mu'}}{\sqrt{\epsilon/\mu} + \sqrt{\epsilon'/\mu'}} = \frac{Z'_0 - Z_0}{Z'_0 + Z_0}. \tag{3.4}$$

The ratio of reflected H to incident H, except for the change of sign, is the same; thus the ratio of reflected power to incident power, or the reflection coefficient for power, is the square of the quantity given in (3.4).

In most common cases, the media are nonmagnetic, and

$$\mu = \mu' = \mu_0,$$

the value characteristic of free space. In this case, from (1.9), Chap. VIII, the square roots in (3.4) are proportional to the corresponding indices of refraction, and the reflection coefficient for power is

$$\text{Reflection coefficient for power} = \frac{(n - n')^2}{(n + n')^2}. \tag{3.5}$$

We note that, if n and n' are interchanged, or the incident wave is approaching the surface from the other side, the reflection coefficient is unchanged; a surface always reflects equally in both directions. The reflection coefficient is of course less than unity. We can easily find E'_x, the amplitude of the refracted wave, if we choose, and show that the power carried to the surface by the incident wave is split up between the reflected and refracted waves, so that, if any energy is in the refracted wave, the reflected wave must be weaker than the incident, and the reflection coefficient must be less than unity. For instance, for a wave of light reflected at a surface between glass and air, we have $n = 1$ for air, and about 1.5 for glass; in this case the coefficient is $(0.5)^2/(2.5)^2 = 1/25$, showing that only 4 per cent of the intensity is reflected from a glass plate at normal incidence.

4. Fresnel's Equations.—Next we take the case of an arbitrary angle of incidence. Here we meet the question of polarization. The vector **E** is at right angles to the direction of propagation, but that does not fix the direction uniquely, and it is said that the wave is polarized in a particular direction if its electric vector (or, according to an alternative convention, its magnetic vector) points in that direction. Let us then consider the two extreme cases. We take the wave

normal of the incident wave to be in the xz plane, as before. Then we consider the case where the electric vector is along the y axis, and the case where it is in the xz plane, as in Fig. 25. In case 1, the y axis points down into the paper, **E** and **E**$'$ point down, and **E**$''$ points up; in case 2, **H**, **H**$'$, and **H**$''$ all point down, along the y axis. We now discuss these cases separately.

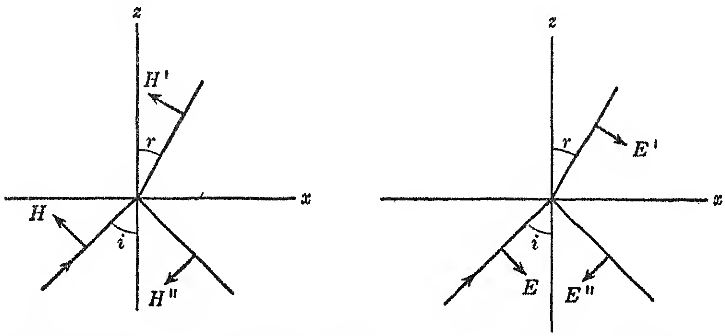


FIG. 25.—Vectors in reflection and refraction. Case 1: y axis points down into the paper. E and E' point down, E'' points up. Case 2: H, H', H'' all point down.

Case 1.—The electric vector is along the y axis, or at right angles to the plane of incidence. All vectors depend on space in the way indicated by (2.2) for the incident wave, (2.3) for the refracted wave, and (2.5) for the reflected wave. From the figure, we see that for the incident wave $H_x = -(E_y/Z_0)\cos i$, where Z_0 is given in (3.1). Similarly for the refracted wave, $H'_x = -(E'_y/Z'_0)\cos r$, and for the reflected wave $H''_x = -(E''_y/Z_0)\cos i$. Hence we have the following relations:

Tangential component of **E**: $E_y - E''_y = E'_y$.

Tangential component of **H**: $-(E_y + E''_y)\dfrac{\cos i}{Z_0} = -E'_y\dfrac{\cos r}{Z'_0}$.

Remembering Snell's law, the second may be rewritten

$$E_y + E''_y = E'_y\frac{\mu}{\mu'}\frac{\tan i}{\tan r}.$$

From this at once, multiplying the first by $(\mu/\mu')(\tan i/\tan r)$, and subtracting, we have

$$E_y\left(\frac{\mu}{\mu'}\frac{\tan i}{\tan r} - 1\right) = E''_y\left(\frac{\mu}{\mu'}\frac{\tan i}{\tan r} + 1\right),$$

$$\frac{E''_y}{E_y} = \frac{\mu\tan i - \mu'\tan r}{\mu\tan i + \mu'\tan r}.$$

If both media are nonmagnetic, so that $\mu = \mu'$, as is usually the case, this becomes

$$\frac{E_y''}{E_y} = \frac{\tan i - \tan r}{\tan i + \tan r} = \frac{\sin i \cos r - \cos i \sin r}{\sin i \cos r + \cos i \sin r},$$

$$\frac{E_y''}{E_y} = \frac{\sin (i - r)}{\sin (i + r)}. \tag{4.1}$$

This gives the amplitude of the reflected wave, and is one of Fresnel's equations. We note that, as i and r become zero, the law of refraction becomes $i/r = n'/n$, $i = (n'/n)r$. Thus, in the limit of normal incidence, the ratio approaches $(n' - n)/(n' + n)$, as we found above (when we correct the signs for the different convention used here). We also note, in the other extreme of tangential or grazing incidence, that $i = 90°$, so that the ratio is $[\sin (90° - r)]/[\sin (90° + r)] = 1$. That is, the reflection coefficient equals unity for grazing incidence. The formula gives a monotonic increase of amplitude as the angle of incidence increases.

Case 2.—The electric vector is in the xz plane, or in the plane of incidence. We let **H** be along the y axis in all waves, and let E, E', E'' be the magnitudes of **E** in the three waves. Then $H_y = E/Z_0$, $H_y' = E'/Z_0'$, $H_y'' = E''/Z_0$. For the components of the electric vectors, we have $E_x = E \cos i$, $E_x' = E' \cos r$, $E_x'' = -E'' \cos i$. Then we have

$$\text{Tangential component of } \mathbf{E}: (E - E'') \cos i = E' \cos r$$

$$\text{Tangential component of } \mathbf{H}: \frac{(E + E'')}{Z_0} = \frac{E'}{Z_0'}.$$

We may rewrite the first as $E - E'' = E' (\cos r/\cos i)$, the second as, $E + E'' = (Z_0/Z_0')E'$. Multiplying the first by

$$\frac{Z_0}{Z_0'} = \left(\frac{\mu}{\mu'}\right)\left(\frac{n'}{n}\right) = \left(\frac{\mu}{\mu'}\right)\left(\frac{\sin i}{\sin r}\right),$$

the second by $\cos r/\cos i$, and subtracting, we have

$$E\left(\frac{\mu}{\mu'}\frac{\sin i}{\sin r} - \frac{\cos r}{\cos i}\right) = E''\left(\frac{\mu}{\mu'}\frac{\sin i}{\sin r} + \frac{\cos r}{\cos i}\right),$$

$$\frac{E''}{E} = \frac{\mu \sin i \cos i - \mu' \sin r \cos r}{\mu \sin i \cos i + \mu' \sin r \cos r}.$$

If $\mu = \mu'$, we can use the trigonometric relations

$$\begin{aligned}\sin (i \pm r) \cos (i \mp r) &= (\sin i \cos r \pm \cos i \sin r)(\cos i \cos r \pm \sin i \sin r) \\ &= \sin i \cos i(\cos^2 r + \sin^2 r) \pm \sin r \cos r(\sin^2 i + \cos^2 i) \\ &= \sin i \cos i \pm \sin r \cos r.\end{aligned}$$

Hence we have

$$\frac{E''}{E} = \frac{\sin (i - r) \cos (i + r)}{\sin (i + r) \cos (i - r)} = \frac{\tan (i - r)}{\tan (i + r)}. \tag{4.2}$$

This is the other of Fresnel's equations.

There is one interesting feature met in this case, which is not present when the electric vector is at right angles to the plane of incidence: if $i + r = 90°$, a perfectly possible situation, the denominator of (4.2) is infinite, so that the reflection coefficient is zero. This angle is called the "polarizing angle"; if unpolarized radiation, consisting of a mixture of both types of radiation, falls on a surface at this angle, only the radiation with its electric vector at right angles to the plane of incidence will be reflected, and the reflected radiation will be polarized. It was by this phenomenon that polarized light was first discovered. Light was reflected from one mirror at this angle. Then its polarization was found by reflecting from a second mirror at the same angle. As the second mirror was rotated about the beam as an axis, so that the polarization changed from being at right angles to the plane of incidence to being in the plane, the doubly reflected beam changed from a maximum intensity to zero. The polarizing angle r' is fixed by $i' + r' = 90°$, and this occurs when $\cos i' = \sin r'$. Using the law of refraction, we find

$$\tan i' = \frac{n'}{n},$$

thus fixing the polarizing angle i'.

5. Total Reflection.—For radiation passing from a dense medium to a rarer medium, so that $n' < n$, the angle of refraction becomes $90°$ for an angle i for which $\sin i = n'/n$. For greater angles of incidence, the equation $\sin r = (n/n') \sin i$ would indicate that $\sin r$ should be greater than unity, so that r must be imaginary. To find the physical meaning of this situation, we compute $\cos r$, which is given by

$$\cos r = \pm \sqrt{1 - \sin^2 r} = \pm j \sqrt{(n/n')^2 \sin^2 i - 1}.$$

The expression (2.3) for the disturbance in the rarer medium then becomes

$$e^{j\omega[t - (x \sin r/v')]}e^{-\omega[\sqrt{(n/n')^2 \sin^2 i - 1}]z/v'},$$

where we have used the negative square root. The first term represents a wave propagated along the x axis, or parallel to the surface of the medium, with an apparent velocity $v'/\sin r$, a value less than v'. The second factor indicates that the amplitude of this wave is damped out as z increases, or as we go away from the surface, so that the wave fronts (surfaces of constant phase) are at right angles to the surfaces of constant amplitude. This disturbance ordinarily damps out in a very short distance. Thus if $(n/n') \sin^2 i$ is decidedly greater than 1, the exponential becomes small when z is a few wave lengths ($\omega z/v'$ a reasonably large number). Consequently the disturbance in the rare medium is not observed, unless special experiments are devised to find it. It is easily shown that Poynting's vector for this wave has no component normal to the surface, so that it does not carry any energy away. Thus all the energy in the incident wave must reappear in the reflected wave, and for this reason the phenomenon is called "total reflection." The angle of incidence given by $\sin i = n'/n$, which must be exceeded in order to have total reflection, is called the "critical angle."

Although there is no change of amplitude on reflection in this case, there is a change of phase, which is sometimes of interest, and which may be treated by Fresnel's equations. Thus in case 1 we have, assuming nonmagnetic media,

$$\frac{E''}{E} = \frac{\sin i \cos r - \cos i \sin r}{\sin i \cos r + \cos i \sin r} = \frac{\sqrt{\sin^2 i - (n'/n)^2} - j \cos i}{\sqrt{\sin^2 i - (n'/n)^2} + j \cos i}$$

and in case 2,

$$\frac{E''}{E} = -\frac{(n/n')^2 \sqrt{\sin^2 i - (n'/n)^2} - j \cos i}{(n/n')^2 \sqrt{\sin^2 i - (n'/n)^2} + j \cos i}$$

In each case, E''/E is the ratio of a complex number to its complex conjugate, which is therefore a complex number of unit magnitude (showing that the reflection coefficient is unity), but with a certain phase angle. Thus, in the general case, where $\mathbf{E}$ has components both in the xz plane and along the y axis, there is a difference of phase between these components upon total reflection, and linearly polarized light in general will become elliptically polarized upon total reflection. To see this, we note that two vibrations at right angles, with the same frequency and phase, produce a resultant vector whose extremity moves in a line (plane polarization), but if the two components are in different phases the extremity of the vector traces out an ellipse. If the phases differ by 90°, and the amplitudes are equal, the polarization is circular.

6. Damped Plane Waves, Normal Incidence.—So far, we have taken up only the case of nonabsorbing media, with real indices of refraction. The case of reflection by an absorbing medium, such as a metal, can be handled, as we saw in Chap. IX, by replacing the index of refraction n by the complex quantity $n - jk$. Considering the case of normal incidence, using (3.4), and setting $\mu = \mu'$, we have

$$\frac{E_x''}{E_x} = \frac{n - (n' - jk)}{n + (n' - jk')} = \frac{(n - n') + jk'}{(n + n') - jk'}, \tag{6.1}$$

on the assumption that the wave is incident on an absorbing medium (n',k') from a nonabsorbing medium. Since this ratio is complex, we see in the first place that there is a phase change on reflection from an absorbing medium. To find the reflection coefficient R, we must take the square of the magnitude of the quantity in (6.1). This can be done by multiplying by the complex conjugate. We have

$$R = \frac{(n - n')^2 + k'^2}{(n + n')^2 + k'^2}.$$

It is interesting to consider the behavior of a metal at relatively low frequencies. We saw in Chap. IX, Sec. 3, that at frequencies below the shorter part of the infrared, n' and k' both approach $\sqrt{\sigma/2\omega\epsilon_0}$. Furthermore, both these quantities are very large compared with unity, or with n, the index of refraction of the transparent medium in question. Substituting these values, the reflection coefficient becomes

$$R = 1 - 2n\sqrt{\frac{2\omega\epsilon_0}{\sigma}}. \tag{6.2}$$

According to our assumptions, the second term is small; thus nearly all the incident radiation is reflected. Furthermore, the second term becomes smaller as the frequency is reduced, or as the conductivity increases. In these cases, very little power flows into the conductor, and is dissipated in it because of the damping of the wave, or, in physical language, because of the dissipation of heat in the resistance. The reason so little power flows into the metal is easily seen, from the relation (2.1) of Chap. VIII for the ratio of E to H. From that equation, in the limit of low frequency, we have

$$\frac{E_x}{H_y} = \sqrt{\frac{j\mu\omega}{\sigma}} = \sqrt{\frac{\mu_0}{\epsilon_0}}\sqrt{\frac{j\omega\epsilon_0}{\sigma}},$$

if we assume that $\mu = \mu_0$. That is, the ratio of E to H is much smaller than the value $\sqrt{\mu_0/\epsilon_0}$ characteristic of empty space. It is so small, in fact, that the E within the metal can almost be neglected. Thus

Poynting's vector at the surface is very small, representing a small energy flow into the metal. The situation is almost like that with a perfect conductor, in which the electric vectors of the incident and reflected waves exactly cancel at the surface of the metal, whereas the magnetic vectors are equal in magnitude, because of the perfect reflection, and add.

Looking back to the value of γ for a metal, from Eq. (1.5) of Chap. VIII, we see that in our limiting case

$$\gamma = \pm j\omega \sqrt{-\frac{j\sigma\mu}{\omega}} = \pm j \sqrt{-j\omega\sigma\mu}.$$

This tells us in the first place that the real and imaginary parts are equal [since $j\sqrt{-j} = (1 + j)/\sqrt{2}$], so that a wave inside a good conductor damps down to a small fraction of its intensity in a few wave lengths; and that the distance in which the intensity falls off to a fraction of its value at the surface decreases as either the frequency or conductivity increases. It is often convenient to define the distance in which the amplitude falls to $1/e$ of its value as the skin depth, for this damping of the wave inside the conductor is simply another way of describing the skin effect, familiar in high-frequency electric-circuit theory. We have for δ, the skin depth,

$$\delta = \sqrt{\frac{2}{\omega\sigma\mu}}. \tag{6.3}$$

For the limiting case of a perfect conductor, we then see that the reflection coefficient is unity; the tangential E at the surface is zero (as it must be, for otherwise the current would have to be infinite), but the tangential H is finite; but that the field within the conductor is a damped wave that falls off infinitely rapidly to zero as we penetrate the metal, so that directly below the surface H, as well as E, is zero. In other words, to take account of the rapid decrease of the tangential H from a finite value at the surface to zero directly below the surface, Stokes's theorem shows that we must assume a surface-current density at the surface of the conductor, numerically equal to the tangential H, but at right angles to it, which because of the perfect conductivity can flow without a corresponding tangential electric field. These limiting boundary conditions are often convenient to use directly in discussing the boundary conditions of an electromagnetic wave reflected by a perfect conductor, as for instance in considering propagation in wave guides, and electromagnetic fields in resonant microwave cavities.

7. Damped Plane Waves, Oblique Incidence.—If a plane wave approaches the surface between a transparent medium and an absorbing medium, such as a metal, at oblique incidence, the problem is considerably more complicated than any that we have taken up so far, and we shall not give a complete discussion. (A more complete treatment is given in J. C. Slater, *Microwave Transmission*, Sec. 13, McGraw-Hill Book Company, Inc., New York, 1943.) We can see the reason for the complication in a very simple way. The wave normal in the metal will be at an angle of refraction r, which certainly will not bring it along the z axis, or the normal to the surface. Thus the surfaces of constant phase, or the wave fronts, in the metal, will not be parallel to the planes z = constant. On the other hand, the amplitude is constant over the surface $z = 0$, the surface of the metal; thus it will also be constant at surfaces z = constant within the metal, since at all points of such a surface the wave will have penetrated equal distances through the absorbing medium, and will have had equal energy losses. Thus the surfaces of constant phase do not coincide with the surfaces of constant amplitude.

We have already seen, in Sec. 5, an example of a case where these surfaces do not coincide; in that case, in fact, they were at right angles. We may get a general approach to this problem by assuming, in place of (1.4) of Chap. VIII, a general solution of the wave equation of the form $e^{j\omega t - \boldsymbol{\gamma}\cdot\mathbf{r}}$, where $\boldsymbol{\gamma}$ and $\mathbf{r}$ are now vectors, and where $\boldsymbol{\gamma}$ can be called the "propagation vector." In Chap. VIII, we could have used this form, with the assumption that γ pointed along the z axis. We saw in (1.6) of Chap. VIII that γ could be complex; but now we see that an additional complication can arise, in that the real and imaginary parts of $\boldsymbol{\gamma}$, which we can call $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, can be vectors that do not have to have the same direction in space. If we substitute our expression in the wave equation, we find that, if there is no loss in the medium, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ must be at right angles to each other, and there is a definite relation between their magnitudes. This is the situation met in total reflection, where we have a damped wave, but in a nonabsorbing medium. The physical situation is clear: since there is no loss, the intensity of the wave must be constant along the direction of propagation, but there is nothing to prevent its intensity varying at right angles to this direction.

On the other hand, in an absorbing medium, the relation between the directions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is quite arbitrary. Thus we can always choose the direction of $\boldsymbol{\alpha}$ so as to make the surface of the metal a surface of constant amplitude, and yet can choose the direction of $\boldsymbol{\beta}$

so as to satisfy the law of refraction. We find, however, that the law of refraction is no longer the simple form of Snell's law. The angle of refraction acts like a complex angle, and we can modify Snell's law, if we choose, to give the correct results, by taking a complex index of refraction, and complex angle of refraction. The net result of the law of refraction, however, is very simple for a metal of good conductivity. We have seen that its index of refraction n', as well as the absorption coefficient k', are very large. The refracted wave within the metal resembles that in an ordinary case of refraction to the extent that, with a very large index of refraction, the angle of refraction is very small, or the wave normal of the refracted wave is nearly along the normal to the surface. In other words, quite independent of the angle of incidence, the wave inside the metal is similar to that discussed in Sec. 6. There is, however, a change of phase on reflection, which we can get from Fresnel's equations, treating the angle of refraction as complex, which is a function of the angle of incidence, and which is different for the two planes of polarization of the incident wave. Thus it can happen that the state of polarization of the incident wave can be changed by reflection, in a characteristic manner, and this gives one of the practical experimental methods for finding information regarding the optical constants of a metal.

Problems

1. Light is reflected from glass of index of refraction 1.5. Compute and plot curves for the reflected intensity as a function of angle, for both sorts of plane polarization.

2. Find the intensity of light in the refracted medium, for arbitrary angle of incidence and both types of polarization. Show that the amount of energy striking the surface is just equal to the amount carried away from it. Note that the amount striking the surface is computed, not from the whole of Poynting's vector, but from its normal component.

3. Light passes normally through a glass plate. Find the weakening in intensity because of the reflection at the faces.

4. Ten plates of glass of index 1.5 are placed together and used as a polarizer. Light strikes the plates at the polarizing angle, and the transmitted light is used. Since all the reflected light is of one polarization, and the reflection at both surfaces of all plates is enough to remove practically all the light of this polarization, the transmitted light will be practically polarized in the other direction. Find the intensity of both sorts of light in the transmitted beam, assuming initially unpolarized light, and hence show how much polarization is introduced. You may have to consider multiple internal reflection.

5. The resistivity of copper is about 1.7×10^{-6} ohm-cm. Calculate the reflective power of copper for wave lengths of light $\lambda = 12\mu$ and $\lambda = 25.5\mu$. The observed values of $1 - R$ are 1.6 per cent and 1.17 per cent at these wave lengths.

CHAPTER XI

WAVE GUIDES AND CAVITY RESONATORS

If electromagnetic radiation is introduced into the open end of a hollow pipe bounded by perfectly conducting, and hence perfectly reflecting walls, the radiation will be reflected from the wall whenever it strikes it, and will be able to progress down the pipe for an indefinite distance. Such a hollow pipe is called a "wave guide," and has recently come into much prominence in connection with the applications of microwaves, or electromagnetic waves whose lengths are of the order of magnitude of laboratory dimensions. The propagation of waves down a wave guide is not quite so simple as it seems at first sight; for the various reflected waves interfere with each other, in such a way that only certain types of waves can exist in the guide. In particular, the field pattern of the wave in a plane perpendicular to the axis can take on only one of a discrete, though infinite, number of characteristic patterns, called "modes," determined by the shape and size of the cross section. Corresponding to each of these modes, a wave is possible in the guide for any given frequency, and the effective wave ength as measured along the guide, the so-called "guide wave length," is determined in terms of the frequency.

For each frequency and guide wave length we can at once determine the phase velocity of propagation along the axis of the guide, and we find that in most cases this phase velocity is greater than the velocity of light, a fact that does not contradict the principle of relativity, since it can be shown that no signal can be propagated with this phase velocity. The phase velocity varies with frequency; that is, the propagation shows dispersion, just like the propagation of light in a dispersive medium, which we discussed in Chap. IX. In fact, as the frequency decreases, the guide wave length, and the phase velocity, increase without limit, both becoming infinite at a quite finite frequency. At still lower frequencies, the disturbance, instead of being propagated along the wave guide, is exponentially attenuated or damped, at a more and more rapid rate as the frequency approaches zero. The frequency at which the disturbance changes over from being propagated to being attenuated is called the "cutoff frequency," and the corresponding wave length of a wave in free space is called

the "cutoff wave length." Since the guide will propagate only waves of free-space wave length shorter than the cutoff wave length, or of frequencies higher than the cutoff frequency, it forms a filter, passing only high frequencies. The cutoff frequency is different for each mode. Thus, for frequencies less than the lowest cutoff frequency, no wave can be propagated down the guide; between the lowest and next lowest cutoff frequencies, only one mode can be propagated; and so on. As a practical matter, wave guides are usually used in the region between the lowest and next lowest cutoff frequencies, so as to have only one mode propagated, the rest attenuated. For certain special types of wave guides, the lowest cutoff frequency is zero, so that in such cases any frequency, no matter how low, can be propagated. We shall find that in such a case this mode of zero cutoff frequency, called the "dominant mode," has the interesting property that the velocity of propagation is equal to the velocity of light in free space, so that this mode, unlike the others, does not show dispersion.

If a length of wave guide is closed by two perfectly reflecting planes, radiation propagated down the guide will be reflected back and forth by the planes, and will form standing waves. Furthermore, in order that the successive reflections may reinforce each other, it is clearly necessary that the round trip, from one plane to the other and back again, should occupy a whole number of periods of the vibration, or that this distance should be a whole number of guide wave lengths. Thus the guide wave lengths, and hence the frequencies, are determined in terms of the dimensions. Such a guide with closed ends forms a simple example of a cavity resonator. It is found, in fact, that any closed region bounded by perfectly reflecting walls can vibrate in certain normal modes, each of a definite frequency fixed by the geometry of the cavity, there being an infinite number of such normal modes of greater and greater frequency. Such resonant cavities form the equivalent, in the microwave range of frequencies, of resonant circuits in the ordinary range where electric circuits are used. It is obvious that wave guides with closed ends, forming resonant cavities, form a close analogy to organ pipes, and it is not surprising to find that Lord Rayleigh, some of whose best known work was in the theory of sound, worked out the theory of wave guides and of cavity resonators in the latter part of the nineteenth century. Their practical application, however, has waited until the last few years, when techniques for producing the oscillations of the required very high frequency have become available.

The properties of wave guides, which we have outlined above,

seem at first sight strange and unexpected, though it is very easy to prove these properties by setting up appropriate solutions of Maxwell's equations. A little further consideration, however, shows that the properties of wave guides follow largely from the discussion of reflection which we have already given. Even the attenuated waves for frequencies below cutoff prove to have their simple explanation, and to be closely related to the attenuated wave in the rarer medium in the problem of total internal reflection, the cutoff phenomenon having an analogy to the critical angle. Accordingly, so as to make the subject physically clear, we shall start our discussion by considering a very simple problem, the propagation of radiation between two parallel plane mirrors, which shows many of the properties met with wave guides. We shall not push this method too far, however, for in fact it is only with rectangular wave guides that one can build up the solution from a study of reflection by plane walls. In more complicated cases, such as the circular guide, this method is not available, and yet we can set up a general treatment that takes care of any arbitrary cross section in a fairly simple way.

1. Propagation between Two Parallel Mirrors.—A perfect reflector allows no field to penetrate within it; we have seen how this is to be realized in fact, by considering a good conductor, in which the field penetrates only to a distance comparable with the skin depth, which reduces to zero in the limit of perfect conductivity, which we assume. Thus the electric and magnetic fields are zero within the mirror, and our boundary conditions reduce to two statements: first, that the tangential component of **E** at the surface of the mirror is zero, since it must be zero within the mirror, and we know that the tangential component is continuous; secondly, the normal component of **B** at the surface must be zero, since **B** is zero within the mirror, and the normal component of **B** is continuous. The other two conditions, on the normal component of **D** and the tangential component of **H**, lead to information about the surface charge and current; for our surface is conducting and can carry charge and current, so that the value of the normal component of **D** outside the metal measures the surface-charge density, and the tangential component of **H** the surface-current density, which appear on the surface. Because of the lack of penetration of the field into the mirror, we need consider only incident waves striking the mirror, and corresponding reflected waves, but no refracted waves within the mirror.

With this understanding of the nature of the reflection process, let us consider a wave striking a perfectly reflecting mirror, and the

corresponding reflected wave. Let the mirror be in the xz plane, the plane $y = 0$, and let the incident wave have its normal in the yz plane, making an angle of incidence i with the y axis, as shown in Fig. 26. Let the incident wave approach from positive values of y, and negative z's. Then, using the same sort of arguments as in Chap. X, the exponential factor representing wave propagation in the incident wave will be

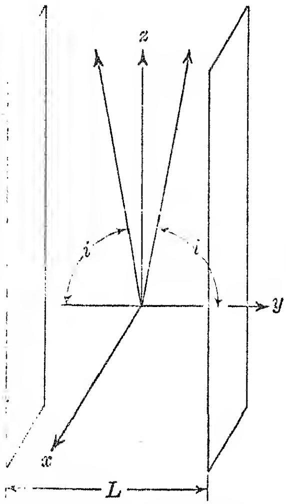


FIG. 26.—Coordinate system for propagation between parallel planes.

$$e^{j\omega\{t-[(-y\cos i+z\sin i)/c]\}} \tag{1.1}$$

and the corresponding factor in the reflected wave will be

$$e^{j\omega\{t-[(y\cos i+z\sin i)/c]\}}. \tag{1.2}$$

We must now superpose these in such a way that certain components of the field (the tangential component of **E**, and the normal component of **B**) will be zero on the plane $y = 0$. The only way in which we can accomplish this is to superpose (1.1) and (1.2) in such a way that the function of y turns into a sine function. We do this by subtracting (1.2) from (1.1), and dividing by $2j$. Thus we obtain the function

$$e^{j\omega[t-(z\sin i)/c]}\sin\left(\frac{\omega\cos i}{c}\right)y. \tag{1.3}$$

The solution (1.3) may be interpreted as a disturbance that is propagated along the z axis with an apparent velocity $c/\sin i$, which is greater than c, since $\sin i$ is less than unity; and varying sinusoidally along y, or corresponding to standing waves along y, with an effective wave length given by

$$\frac{2\pi}{\lambda_c} = \frac{\omega\cos i}{c}, \qquad \lambda_c = \frac{\lambda_0}{\cos i}, \tag{1.4}$$

where λ_c is the effective wave length along y, $\lambda_0 = 2\pi c/\omega$ is the free-space wave length, or wave length of the disturbance in the absence of reflections. The standing-wave phenomenon along y can be observed with visible light, in the form of the so-called "Lippmann fringes." Lippmann showed that, if standing waves are set up by this process, not in free space, but in the sensitive emulsion of a photo-

graphic plate, the emulsion will be blackened along surfaces half a wave length apart in the standing-wave pattern. The phenomenon can also be demonstrated very easily, with the longer waves of the microwave region.

We have so far considered only the reflection by one mirror, at $y = 0$. In case, however, we have a second mirror, at $y = L$, we must satisfy the same sort of boundary conditions at this second mirror. Thus our function (1.3) must be zero, not only at $y = 0$, but also at $y = L$. This leads at once to the condition

$$L = \frac{n\lambda_c}{2}, \tag{1.5}$$

where n is an integer. In other words, the dependence of the disturbance on y cannot have any arbitrary wave length, but only an infinite set of discrete wave lengths. These form the modes that we have already mentioned. It should be noted that, to discuss this problem of two mirrors, we do not have to think of multiple reflection, in the elementary sense. Our one wave traveling along $+y$, and the other traveling along $-y$, suffice to satisfy the boundary conditions everywhere. If we treat the disturbance by means of rays, we should say that any given ray is reflected back and forth an infinite number of times from the two mirrors, but in fact all the rays traveling to the right coalesce to furnish the description of the single wave traveling to the right, and similarly with the rays traveling to the left.

For a given n, or a given mode, the wave length λ_c is determined by (1.5). This is independent of the free-space wave length λ_0, or the corresponding frequency ω. Knowing λ_0, from the conditions of our problem, we can next use (1.4) to find i, the angle of incidence, and from this we can find the apparent phase velocity v_g along the z direction, determined by

$$v_g = \frac{c}{\sin i}, \tag{1.6}$$

and the corresponding wave length,

$$\lambda_g = \frac{v_g}{c}\lambda_0 = \frac{\lambda_0}{\sin i}. \tag{1.7}$$

For many purposes it is convenient to eliminate i, the angle of incidence. Computing $\cos i$ from (1.4), $\sin i$ from (1.7), and taking the sum of their squares, we find at once the relation

$$\frac{1}{\lambda_g^2} + \frac{1}{\lambda_c^2} = \frac{1}{\lambda_0^2}, \tag{1.8}$$

This simple equation determines the wave length λ_g along the z direction, in terms of the free-space wave length λ_0, and the wave length λ_c along the y direction, or the direction normal to z. We notice one interesting fact from (1.8). If λ_c is fixed, and if λ_0 is increased, or the frequency is decreased, $1/\lambda_g^2$ decreases, or λ_g increases, with consequent increase of the phase velocity v_g, until finally when λ_0 equals λ_c, λ_g and v_g become infinite. For values of λ_0 greater than λ_c, the wave length λ_g and the velocity v_g become imaginary, so that the wave is attenuated rather than propagated along the z axis. In other words, we have the phenomena that we have already described, including a cutoff wave length, with only attenuation for longer waves than the cutoff wave length. Furthermore, we see that λ_c is the cutoff wave length in question.

We have so far considered only the exponential function associated with the direct and reflected waves, and the corresponding relations regarding the wave length and frequency. In addition, we should take up the orientation and magnitude of the electric and magnetic vectors. As in Chap. X, there are two cases: that in which **E** is along the x direction, or at right angles to z for both the direct and reflected waves, and that in which **H** is in the x direction. In the language of wave guides, these are known, respectively, as the transverse electric (abbreviated *TE*) and transverse magnetic (abbreviated *TM*) modes of propagation. All wave guides, we shall find, have modes of both types. We could go through the equivalent of Fresnel's equations, and find the relations between **E** and **H** in these two types of modes, but it is so easy to prove much more general results holding for any type of guide, as we shall do in the next section, that we shall not carry out this discussion here.

Our problem of propagation between two mirrors forms a very simple and yet instructive case of wave-guide propagation. If we add two more reflecting surfaces perpendicular to the x axis, we have a rectangular wave guide. Some of the modes encountered with the mirrors, namely, those in which **E** is along the x axis, satisfy the boundary conditions for the rectangular guide as well, and in fact form important modes of this guide. They are not the whole set of modes, however, for there can be a sinusoidal variation along x as well as along y. If we furthermore add two reflecting surfaces perpendicular to z, we form a totally enclosed rectangular cavity We now cannot use the solution we have so far written, involving propagation along z; we must rather superpose a reflected wave traveling along $-z$, giving a standing wave along z. To satisfy the boundary

conditions at the mirrors, which may be located for example at $z = 0$, $z = M$, we must make the sinusoidal function along z, whose wave length is λ_g, satisfy a condition analogous to (1.5), or

$$M = \frac{m\lambda_g}{2}, \tag{1.9}$$

where m is an integer. Combining (1.5), (1.9), and (1.8), this leads to

$$\left(\frac{m}{2M}\right)^2 + \left(\frac{n}{2L}\right)^2 = \frac{1}{\lambda_0^2}. \tag{1.10}$$

That is, the free-space wave length, or the frequency, of the disturbance in the cavity is determined by the integers n and m. We have a discrete set of normal modes of oscillation, in which the cavity can resonate.

The problem is similar to that of the elastic vibrations of a coupled system or of a vibrating membrane in mechanics, in which there are a discrete set of normal modes of oscillation, each with its own resonant frequency, and characteristic wave form. Furthermore, as in mechanics, we can introduce normal coordinates describing the various modes, using these normal coordinates to describe forced oscillations of the system, and we can prove orthogonality relations holding between the various normal vibrations. These subjects are, however, beyond the scope of the present book. We now proceed to a much more general discussion of propagation in wave guides, holding for a wave guide of arbitrary cross section, and reducing to the results of the present section in our present simple case. This general discussion is simpler and more powerful, but it does not exhibit the way in which the field is made up of the superposition of direct and reflected waves.

2. Electromagnetic Field in the Wave Guide.—We shall assume a wave guide in the form of a hollow pipe of arbitrary cross section in the xy plane, extending indefinitely along z. As shown in Eqs. (1.2), Chap. VIII, each of the components of **E** and **H** must satisfy the wave equation, which, in empty space, is

$$\nabla^2\mathbf{E} - \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} = 0, \qquad \nabla^2\mathbf{H} - \frac{1}{c^2}\frac{\partial^2\mathbf{H}}{\partial t^2} = 0. \tag{2.1}$$

We now ask if we can obtain a solution of the form $u(x,y)e^{j[\omega t-(2\pi z/\lambda_g)]}$, where $u(x,y)$ is a function of x and y, and where λ_g, the guide wave length, has the same significance as in Sec. 1. Substituting in (2.1), and using $\lambda_0 = 2\pi c/\omega$, we find that u satisfies the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \left(\frac{2\pi}{\lambda_c}\right)^2 u = 0, \tag{2.2}$$

where λ_c is defined in terms of λ_g and λ_0 by (1.8). Equation (2.2) must be solved subject to certain boundary conditions, which we shall discuss presently. Subject to these boundary conditions, we find solutions for only a discrete set of values λ_c. From (1.8) we are then led, as in the preceding section, to the conclusion that λ_c forms a cutoff wave length, such that free-space wave lengths greater than λ_c are attenuated in the mode in question, while shorter free-space wave lengths are propagated. Thus those conclusions are quite general, and are not limited to the case of propagation between parallel reflecting planes.

The fields **E** and **H** must satisfy not only the wave equation, but Maxwell's equations as well. When we write these equations down, we find at once, as mentioned in the preceding section, that two separate types of solutions are possible: solutions for which $E_z = 0$, the transverse electric, or *TE*, waves, and solutions for which $H_z = 0$, the transverse magnetic, or *TM*, waves. We shall let $\mathbf{E}_t$, $\mathbf{H}_t$ be the transverse components of **E** and **H** (that is, $\mathbf{E}_t = \mathbf{i}E_x + \mathbf{j}E_y$, and so on). Letting **k** be unit vector along the z axis, as usual, we then find easily and directly from Maxwell's equations the following relations:

$$TE\text{: grad } H_z = 2\pi j \frac{\lambda_g}{\lambda_c^2} \mathbf{H}_t,$$

$$TM\text{: grad } E_z = 2\pi j \frac{\lambda_g}{\lambda_c^2} \mathbf{E}_t. \tag{2.3}$$

In these expressions, H_z and E_z represent that part of the corresponding quantity which must be multiplied by the exponential $e^{j[\omega t - (2\pi z/\lambda_g)]}$ to give the complete components. Since they are functions of x and y only, their gradients are in the xy plane, as is proper for $\mathbf{H}_t$ or $\mathbf{E}_t$. We find that there is a relationship between the transverse components of **E** and **H**, as follows:

$$\mathbf{H}_t = \frac{\mathbf{k} \times \mathbf{E}_t}{Z_0}, \text{ where } Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{\lambda_g}{\lambda_0} \text{ for } TE$$

$$= \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{\lambda_0}{\lambda_g} \text{ for } TM. \tag{2.4}$$

This means, since **k** and $\mathbf{E}_t$ are at right angles to each other, that $\mathbf{H}_t$ is at right angles to $\mathbf{E}_t$ in the xy plane, and is equal in magnitude to the magnitude of $\mathbf{E}_t$, divided by the quantity Z_0, which is clearly analogous to the corresponding quantity introduced in Eq. (2.1), Chap. VIII.

Using (2.3) and (2.4), we can find all the components of **E** and **H**

from H_z (in the TE case) or from E_z (in the TM case). These quantities, like the transverse components, satisfy the wave equation (2.2), and are scalar solutions of that differential equation. Furthermore, E_z is zero on the boundary of the guide (since $\mathbf{E}$ must have no tangential component on the surface), while H_z has a vanishing normal derivative on the boundary (since $\mathbf{H}$ must have no normal component on the surface, and $\mathbf{H}_t$ is proportional to the gradient of H_z). We may draw lines of H_z = constant (in the TE case) or of E_z = constant (in the TM case) in the xy plane. Then by (2.3) the orthogonal trajectories of these lines will be along the direction of $\mathbf{H}_t$ (in the TE case) or of $\mathbf{E}_t$ (in the TM case). Finally, since by (2.4) the direction of $\mathbf{H}_t$ is perpendicular to that of $\mathbf{E}_t$, the lines of constant H_z will be along the direction of $\mathbf{E}_t$ (in the TE case), and the lines of constant E_z will be along the direction of $\mathbf{H}_t$ (in the TM case). Proceeding in this way, we may draw lines of force, for the transverse components of $\mathbf{E}$ and $\mathbf{H}$, in the xy plane, finding of course that the electric lines of force meet the surfaces at right angles, while the magnetic lines of force are tangential to the surface.

For every scalar solution of the two-dimensional wave equation satisfying the condition that it vanishes on the boundary, we get a TM wave, and for every solution whose normal derivative vanishes we get a TE wave, as we have seen above. There will be an infinite number of solutions of each type, each corresponding to a particular cutoff wave length. These wave lengths may be arranged in order of decreasing magnitude; they start with a largest cutoff wave length, associated with the lowest mode of oscillation, and extend indefinitely toward shorter and shorter wave lengths, so that we have an infinite number of modes of oscillation. We shall give in the next section some simple examples of the modes in various types of wave guides.

In some cases, the first mode has an infinite cutoff wave length; in this case we call it a "principal mode." When a mode of infinite cutoff wave length, or principal mode, exists, it has great practical importance, because it can be used to propagate any wave length, no matter how long. The commonly used mode of the coaxial line is a principal mode, and the familiar parallel-wire transmission line, ordinarily used for low frequencies, can also be considered as a wave guide with a principal mode. We find that such a principal mode exists only if the wall of the wave guide consists of at least two separated conductors, as for instance in the coaxial line. The physical reason for this is quite clear: we can put a very low frequency, or direct current, into a transmission line consisting of two or more conductors,

and they will be insulated from each other, and suited to conduct the current. If there is only one conductor, however, as in an ordinary hollow pipe, there would clearly be a short circuit for a low frequency or direct current, and no propagation is possible until we get to a wave length short enough so that something like real wave propagation occurs. For a principal mode, the cutoff wave length is infinite, so that (2.3) tells us that grad H_z, or grad E_z, must be zero. That is, the longitudinal components of both **E** and **H** are zero, and such a wave is simultaneously transverse electric and transverse magnetic. It is sometimes called a "transverse electromagnetic (*TEM*) wave" for this reason. Furthermore for such a wave, as we see from (1.8), the guide wave length becomes equal to the free-space wave length, so that the velocity of propagation becomes c. Finally, from (2.4), the quantity Z_0 for a principal wave becomes $\sqrt{\mu_0/\epsilon_0}$, which by Eq. (2.1), Chap. VIII, is the ratio of magnitudes of **E** and **H** in a plane wave in free space. Thus a principal wave has many of the properties of a wave in free space.

3. Examples of Wave Guides.—The commonest type of wave guide is that of rectangular cross section, bounded by planes at $x = 0$, $x = a$, $y = 0$, $y = b$. The solution of the wave equation (2.2) in this case can be carried out in rectangular coordinates, and is at once obvious. Thus for a *TE* wave we have

$$H_z = \cos\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\, e^{j[\omega t-(2\pi z/\lambda_g)]}, \tag{3.1}$$

where

$$\frac{1}{\lambda_c^2} = \left(\frac{m}{2a}\right)^2 + \left(\frac{n}{2b}\right)^2, \tag{3.2}$$

and where m, n are integers. Using (2.3), (2.4), we then have for the other components

$$\begin{aligned}
H_x &= -\frac{\lambda_c^2}{2\pi j\lambda_g}\frac{m\pi}{a}\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\, e^{j(\omega t-2\pi z/\lambda_g)},\\
H_y &= -\frac{\lambda_c^2}{2\pi j\lambda_g}\frac{n\pi}{b}\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}\, e^{j(\omega t-2\pi z/\lambda_g)},\\
E_x &= -\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{\lambda_c^2}{2\pi j\lambda_0}\frac{n\pi}{b}\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}\, e^{j(\omega t-2\pi z/\lambda_g)},\\
E_y &= \sqrt{\frac{\mu_0}{\epsilon_0}}\frac{\lambda_c^2}{2\pi j\lambda_0}\frac{m\pi}{a}\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\, e^{j(\omega t-2\pi z/\lambda_g)},\\
E_z &= 0.
\end{aligned} \tag{3.3}$$

Of this infinite set of possible *TE* solutions, only one is commonly used in practice, in a rectangular guide. If $a > b$, this is the mode

of longest cutoff wave length, given by $m = 1, n = 0$. We note from (3.3) that for this mode $E_x = 0$, and E_y is proportional to $\sin \pi x/a$, going to a maximum at the center of the guide, and is independent of y, so that the lines of force run straight across the guide. There is no mode corresponding to $m = 0, n = 0$, for then as we see from (3.3) all the transverse components of field are zero, and we verify easily that the field cannot exist. Thus there is no principal mode for the rectangular guide.

For the TM waves, we have

$$E_z = \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} e^{j(\omega t - 2\pi z/\lambda_g)}, \tag{3.4}$$

where the cutoff wave length is again given by (3.2). We thus see that, for the case of the rectangular guide, the TE and TM waves have the same cutoff wave lengths; this is a special case, which does not hold for most shapes of guides. Clearly we must have m and n both equal to or greater than unity, however, in the case of the TM waves, as we see from (3.4), so that the case $m = 1, n = 0$, which provides the dominant TE wave, does not exist for the TM waves. From (3.4) we can find the other components of **E** and **H**, as we did for the TE waves in (3.3).

The next most important example of wave guides after the rectangular guide is that in which the bounding surfaces are circular cylinders. This includes both the circular guide, and the coaxial line, which has two concentric cylinders, with the wave propagated in the annular space between. For either of these cases, we must handle Maxwell's equations and the wave equation in polar coordinates. In Appendix IV, where we discuss vector operations in curvilinear coordinates, we see that the wave equation, equivalent to (2.2), is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \left(\frac{2\pi}{\lambda_c}\right)^2 u = 0. \tag{3.5}$$

This equation can be solved by separation of variables. We assume that u can be written as a product of a function of r, and a function of θ: $u = R(r)\Theta(\theta)$. Substituting, dividing by u, and multiplying by r^2, (3.5) takes the form

$$\frac{r^2}{R}\left[\frac{1}{r}\frac{d}{dr}\left(r\frac{dR}{dr}\right) + \left(\frac{2\pi}{\lambda_c}\right)^2 R\right] + \frac{1}{\Theta}\frac{d^2\Theta}{d\theta^2} = 0. \tag{3.6}$$

The first term of (3.6) is a function of r alone, the second a function of θ alone; hence by the usual argument used in separation of variables,

each term must be a constant. Let us assume then that

$$\frac{1}{\Theta}\frac{d^2\Theta}{d\theta^2} = -n^2, \qquad \frac{d^2\Theta}{d\theta^2} + n^2\Theta = 0, \qquad \Theta = \sin n\theta \text{ or } \cos n\theta. \quad (3.7)$$

Since u must be a single-valued function of position, it is clear that increasing θ by 2π must bring Θ back to its original value; thus the quantity n must be an integer. Substituting the value from (3.7) back in (3.6), and multiplying by R/r^2, we then have as the equation for R

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dR}{dr}\right) + \left[\left(\frac{2\pi}{\lambda_c}\right)^2 - \frac{n^2}{r^2}\right]R = 0 \quad (3.8)$$

Equation (3.8) is Bessel's equation, whose properties are discussed in Appendix VII. It has two independent solutions, called Bessel's function and Neumann's function:

$$R = J_n\left(\frac{2\pi r}{\lambda_c}\right) \quad \text{or} \quad N_n\left(\frac{2\pi r}{\lambda_c}\right). \quad (3.9)$$

These functions have the following properties, discussed in Appendix VII: At $x = 0$, $J_n(x)$ is proportional to x^n, whereas $N_n(x)$ becomes infinite. For large values of x, $J_n(x)$ and $N_n(x)$ approach the values

$$J_n(x) \to \sqrt{\frac{2}{\pi x}}\cos\left(x - \frac{2n+1}{4}\pi\right),$$
$$N_n(x) \to \sqrt{\frac{2}{\pi x}}\sin\left(x - \frac{2n+1}{4}\pi\right). \quad (3.10)$$

From this, it is clear that, in a circular guide, where the field must be finite at $r = 0$, we must use only the Bessel functions J_n. On the other hand, with a coaxial line, in which $r = 0$ does not fall within the region where we are finding the field, the Neumann as well as the Bessel functions can, and in fact must, be used.

We see, then, that the function u which satisfies the wave equation (3.5) must be the form

$$u = \left[AJ_n\left(\frac{2\pi r}{\lambda_c}\right)\cos(n\theta - a) + BN_n\left(\frac{2\pi r}{\lambda_c}\right)\cos(n\theta - b)\right]. \quad (3.11)$$

Here we have introduced amplitude constants A and B, and phase constants a and b. For a *TE* wave, as we have seen before, u stands for H_z, and its normal derivative must be zero at the boundary of the guide; that is, the derivative of the function of r with respect to r

must be zero. For a TM wave, u stands for E_z, which must itself be zero on the boundary. Now let us consider first a circular wave guide, for which we need only the Bessel function J_n. If R_1 is the radius of the guide, we must clearly have $J_n(2\pi R_1/\lambda_c) = 0$ for a TM mode, and $J'_n(2\pi R_1/\lambda_c) = 0$ for a TE mode, where the prime indicates a derivative.

By giving a table of the maxima, minima, and zeros of the Bessel's functions, we can then find the values of λ_c for the various modes. The first few values are given in the following table:

MAXIMA AND ROOTS OF BESSEL'S FUNCTIONS

$J'_n(x_{nm}) = 0(TE)$	$x_{01} = 3.832$	$x_{11} = 1.842$	$x_{21} = 3.05$
	$x_{02} = 7.016$	$x_{12} = 5.330$	$x_{22} = 6.71$
$J_n(x_{nm}) = 0(TM)$	$x_{01} = 2.405$	$x_{11} = 3.832$	$x_{21} = 5.135$
	$x_{02} = 5.520$	$x_{12} = 7.016$	$x_{22} = 8.417$

If we let x_{nm} be the mth value of x for which $J'_n(x) = 0$ (for the TE modes) or for which $J_n(x) = 0$ (for the TM modes), we then clearly have the cutoff wave length given by

$$\lambda_c = \frac{2\pi R_1}{x_{nm}}, \tag{3.12}$$

where the x_{nm}'s are given in the table above. Thus we can calculate the cutoff wave lengths of the various modes. Using (3.11), and the relations (2.3) and (2.4), we can then find the various components of the field, where we remember that the gradient must be computed in polar coordinates, as discussed in Appendix IV.

For a coaxial line, we must satisfy boundary conditions, not at one value of r, but at two. This can be done only by combining the Bessel and Neumann functions, and using the extra arbitrary constant gained thereby. Thus, if the radius of the inner conductor is R_1, and of the outer one is R_2, we may combine the Bessel and Neumann functions in such a way as to satisfy the boundary conditions at R_1. From (3.10) it is clear that a suitable combination of these two functions will act like a cosine or sine function with arbitrary phase, which can be chosen to put the zero, or the maximum, at any desired point. This can be done for arbitrary λ_c. Then we can choose λ_c so as to satisfy the boundary condition at R_2 as well. This process is a little involved, and it is not so easy to get the numerical answers to problems as with the circular guide, where we found the cutoff wave lengths explicitly in (3.12). The main interest in the coaxial line, however, is not in the modes that we investigate in this way, but in the principal

mode, whose cutoff wave length is infinite. This can be approached by a limiting process from the solution in terms of Bessel and Neumann functions, but it is much easier to discuss it directly from the wave equation (2.2).

For a principal mode, where λ_c is infinite, the equation (2.2) reduces to Laplace's equation $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$. Thus the problem of finding u becomes mathematically equivalent to that of finding an electrostatic potential in a two-dimensional problem. In our case of circular symmetry, we may take u to be $\ln r$, so that its gradient points in the direction of r, and is a constant divided by r. With a coaxial line, this gradient may be chosen as **E**, with **H** at right angles to it, and we have satisfied our boundary conditions at the surfaces of the conductors. Thus we have, for the principal mode of a coaxial line, the very simple solutions

$$E_r = \frac{1}{r} e^{j(\omega t - 2\pi z/\lambda_0)}, \qquad H_\theta = \sqrt{\frac{\epsilon_0}{\mu_0}} E_r. \tag{3.13}$$

The wave, like all principal modes, is propagated with the velocity of light. The mode, as in all principal modes, is *TEM*, both **E** and **H** being transverse. This is not the only case of principal modes that we can work out simply. For instance, for a parallel-wire transmission line, we may use our earlier solution for the capacity of two parallel cylindrical conductors, worked out in Sec. 2, Chap. II. There we found the potential of two such conductors, a solution of Laplace's equation, reducing to a constant on the surface of each conductor. The gradient of this potential gave the electrostatic field, cutting each conductor at right angles. By the principles we have just stated, this same field is proportional to the value of **E** in the principal mode of the parallel-wire transmission line, and **H** is at right angles to it, so that the equipotentials are also the lines of magnetic force.

4. Standing Waves in Wave Guides.—We have so far considered a wave propagated along the z direction in an infinitely long wave guide, and varying according to an exponential $e^{j(\omega t - 2\pi z/\lambda_0)}$. Now, however, we consider the possibility that the wave may be reflected from an obstacle at the end, sending back a reflected wave. This reflecting obstacle may be a perfect reflector, such as a metallic wall; then the reflected wave will be equal in amplitude to the incident wave, and there will be standing waves in which there are nodes and antinodes, the amplitudes canceling and resulting in zero disturbance at the nodes, with addition of the amplitudes midway between. On the other hand, the obstacle may be only a partial

reflector, allowing some radiation to penetrate it, reflecting the rest. In this case, the reflected wave will have less intensity than the incident wave, and the interference of the two will never completely cancel the disturbance.

If the amplitude of the transverse electric field in the reflected wave is r times as great as in the incident wave, where r is sometimes called the "reflection coefficient," the amplitude of the resulting disturbance will vary from something proportional to $1 + r$, where the two waves add, to something proportional to $1 - r$, where they oppose each other. The ratio of maximum to minimum E, or $(1 + r)/(1 - r)$, is called the "standing-wave ratio" in voltage. We see that it can go from unity, when $r = 0$, and there is no reflection, to infinity, when $r = 1$, and there is perfect reflection. The points along the guide where the resultant transverse electric field is a maximum are called "standing-wave maxima," and those where it is a minimum are "standing-wave minima." For an infinite standing-wave ratio the standing-wave minima become the nodes, and in any case they come a half wave length apart. It is clear that from a measurement of the magnitude of the standing-wave ratio we can find the reflection coefficient r, and from the positions of the standing-wave minima we can find the phase change on reflection.

There is one simple case in which we can calculate the reflection coefficient easily. That is the case in which part of the guide, say for positive z, is filled with a dielectric, while the rest, say for negative z, is empty. Then a wave approaching the surface of separation from negative z's will be partly reflected, partly transmitted, just as a wave approaching a surface of glass from air is partly reflected, as we have seen in Chap. X. In simple cases, we could use Fresnel's equations, as derived in that chapter, to find the reflection coefficient in the present case, but we shall instead proceed directly, showing that the results are in close analogy to those of Chap. X. First we must consider the nature of the field in the part of the guide filled with the dielectric. If we follow through the derivation of Sec. 2, we find that the only change introduced by filling the space with a dielectric is to change ϵ_0 to ϵ. This has the effect first of changing the velocity of propagation. Thus the free-space wave length λ_0 is equal to $1/n$ times the value for empty space, where n is the index of refraction. For a given cutoff wave length, (1.8) then shows that λ_g will be smaller than in free space, though not in the same ratio as λ_0. In Eq. (2.2), the wave equation for u, we find that there is no change introduced as a result of the dielectric. Thus the problem of finding cutoff wave lengths

and satisfying boundary conditions at the boundary of the guide is as before. In (2.3), the only change in the process of determining the transverse from the longitudinal components of field comes in the changed value of λ_g, which changes the scale of the transverse components, but not the directions of the vectors. Similarly in (2.4) the only change in the relation between transverse $\mathbf{E}$ and $\mathbf{H}$ comes from the change in the quantity Z_0, which is different both because of the direct change of ϵ and because of the change in λ_g.

Now let us consider a direct and a reflected wave for negative z, and a transmitted wave for positive z. If $\mathbf{E}_t^0$ represents the tangential electric field in its dependence on x and y, we may write the tangential electric and magnetic fields in the first medium as

$$\begin{aligned} \mathbf{E}_t &= \mathbf{E}_t^0[e^{j(\omega t - 2\pi z/\lambda_g)} + re^{j(\omega t + 2\pi z/\lambda_g)}] \\ \mathbf{H}_t &= \left(\frac{\mathbf{k} \times \mathbf{E}_t^0}{Z_0}\right)[e^{j(\omega t - 2\pi z/\lambda_g)} - re^{j(\omega t + 2\pi z/\lambda_g)}]. \end{aligned} \tag{4.1}$$

Here the minus sign in the expression for the transverse magnetic field in the reflected wave is necessary so that the Poynting vector of the reflected wave will point along $-z$. Now at $z = 0$, at the boundary of the second medium, we have from (4.1)

$$\mathbf{H}_t = \frac{\mathbf{k} \times \mathbf{E}_t}{Z_0}\frac{1 - r}{1 + r}. \tag{4.2}$$

But, because of the continuity of the tangential components of $\mathbf{E}$ and $\mathbf{H}$, the values of $\mathbf{H}_t$ and $\mathbf{E}_t$ in (4.2) are also the values found in the second medium. On the other hand, by the modification of (2.4) appropriate for the second medium, we must also have

$$\mathbf{H}_t = \frac{\mathbf{k} \times \mathbf{E}_t}{Z_0'}, \tag{4.3}$$

where Z_0' is the value for the second medium. Equating (4.2) and (4.3), we may solve for r, finding

$$r = \frac{Z_0' - Z_0}{Z_0' + Z_0}. \tag{4.4}$$

In (4.4) we have solved the problem of finding the reflection coefficient for the electric field, at a boundary between two media in a wave guide. The analogy of this to Eq. (3.4), Chap. X, is obvious, though the meaning of the symbols is somewhat different. Proceeding in the other direction, we have

$$\frac{1 + r}{1 - r} = \frac{Z_0'}{Z_0}. \tag{4.5}$$

Thus we see that a measurement of reflection coefficient allows us to find the quantity Z'_0 characteristic of the second medium.

Often in actual practice there are other sorts of discontinuities in wave guides in addition to a simple change of index of refraction. For instance, there may be a diaphragm in the guide, with a hole of some arbitrary shape in it. At such an obstacle, part of the wave will be reflected, part transmitted. We can then set up a reflection coefficient, just as we have done in the case we have discussed, and can get the properties of the reflection by measuring the standing-wave ratio and the position of the standing-wave minima. In such cases it is often convenient to define a quantity like the Z'_0 of (4.5), characterizing the obstacle. We have already pointed out, in Sec. 2, Chap. VIII, the analogy between quantities like Z_0 and an impedance; in this way we are allowed to assign certain impedances to obstacles in wave guides, which prove to be pure reactances in case they introduce no losses. A further discussion of such reactances is beyond the scope of this text.

5. Resonant Cavities.—We have already pointed out that, by terminating a guide with two perfectly reflecting walls, we form a resonant cavity. At each such wall, say at $z = 0$, $z = M$, the field will be perfectly reflected, so that there will be a standing-wave pattern set up within the cavity, with infinite standing-wave ratio. To satisfy boundary conditions at both walls, we must have the tangential electric field zero at each, and this can be accomplished only if there is an integral number of half wave lengths in the length M. Thus we have the condition

$$M = \frac{m\lambda_g}{2}, \tag{5.1}$$

where m is an integer. Putting this into Eq. (1.8), we then have

$$\frac{1}{\lambda_0^2} = \frac{1}{\lambda_c^2} + \left(\frac{m}{2M}\right)^2. \tag{5.2}$$

That is, the free-space wave length, and frequency, are definitely determined for a cavity resonance. For each cutoff wave length of the guide, we shall have an infinite number of resonant modes of the resonant cavity; so that, since in general there will be a double infinity of cutoff wave lengths, there will be a triple infinity of resonant modes for the cavity. We can, of course, get formulas for the resonant frequencies in any case in which we can compute the cutoff wave lengths, as in the cases of rectangular and circular guides which we have already discussed.

The resonant cavities formed in this way, by putting reflecting walls at the ends of a guide, form but a special case of the general problem of resonant cavities. Resonant modes of oscillation can exist in any cavity bounded by perfectly reflecting walls, no matter what its shape. In some cases we can solve for the fields (as for instance in the spherical cavity, which we shall mention in the next chapter), but in the great majority of cases there is no practical way to solve the wave equation. We cannot in general reduce the solution of Maxwell's equations to the solution of a scalar wave equation like (2.2), as we have been able to do for the guide; we must find vector solutions from the outset. The problem of discussing fields inside an arbitrary resonant cavity is then a complicated one, but nevertheless we can prove certain general theorems about them, even without being able to solve for them analytically. These theorems are, however, too complicated for us to take up here.

Problems

1. A wave travels along an air-filled coaxial line having an inner conductor of radius a and a sheath of inner radius b, both of negligible resistance, the peak voltage and current of this wave being V and I. Using cylindrical coordinates, with the z axis coincident with the axis of the line, show that the electric and magnetic fields are given by

$$E_r = \frac{Ve^{j(\omega t - \beta z)}}{r \ln b/a}, \qquad H_\theta = \frac{Ie^{j(\omega t - \beta z)}}{2\pi r}$$

What value must β have? Find the characteristic impedance V/I and the wave impedance E_r/H_θ in terms of ϵ_0, μ_0, a, and b.

2. Starting with Eq. (3.4) for the longitudinal electric-field component, derive expressions for the transverse components of **E** and **H** for *TM* waves in a rectangular wave guide.

3. For the case of *TE* and *TM* modes in a rectangular wave guide, compute Poynting's vector, its average value, and from this the power transmitted by these modes. For the lowest *TE* modes in a rectangular wave guide, find an expression for the maximum power that can be transmitted if the electric field cannot exceed a value E_0 without sparking.

4. Find expressions for the field components for *TE* and *TM* waves in a circular wave guide. Discuss the degeneracies present in these modes, especially those resulting from the relation $J_0'(x) = -J_1(x)$.

5. Show by direct substitution in Maxwell's equations that the following field for a *TM* wave is a solution of these equations:

$$H_x = -\frac{k^2}{\beta\omega\mu} E_y = H_0 e^{\pm j\sqrt{k^2-\beta^2}\,y} e^{j(\omega t - \beta z)}$$

$$E_z = \mp \frac{\omega\mu\sqrt{k^2 - \beta^2}}{k^2} H_x,$$

where $k^2 = \omega^2\epsilon\mu - j\omega\mu\sigma$. Use this solution to discuss the propagation of a surface

wave along a metal surface. Let the metal fill the space $y < 0$, with air for $y > 0$. Write the appropriate solutions in both regions, and show from the boundary conditions that β is given by $1/\beta^2 = 1/k^2 + 1/k_0^2$, with $k_0^2 = \omega^2\epsilon\mu$. Now assume that the conductivity σ of the metal is high so that $|k^2| \gg k_0^2$. Compute the total current in the metal per unit width in the x direction, and show that it is equal to the magnetic field at the metal-air surface, just as if one had a perfect conductor. From the y component of the Poynting vector, compute the power flow into the metal at the boundary $y = 0$, and show that it is of the form I^2R, with I the rms total current in the metal. Show further that this resistance R is the same as if the current density were uniform to a depth below the metal surface equal to the skin depth $\sqrt{2/\omega\mu\sigma}$ and zero below this.

6. A wave guide is terminated by a thin conducting surface perpendicular to the axis of the guide. Find the resistance per square of this conductor (that is, the resistance across a square of the material, from one side to the opposite side, which is independent of the size of the square), such that a wave falling on the surface will be totally absorbed, with no reflection. Find the relation of this resistance per square to the quantity Z_0 of (2.4).

7. A wave guide is filled with a dielectric giving a value of Z_0[from (2.4)] for $z < 0$; another dielectric giving Z_1 for $0 < z < L$; and a third giving Z_2 for $L < z$. A wave travels in the direction of increasing z, and in addition there are reflected waves, in the direction of decreasing z, for $z < 0$, and for $0 < z < L$. By considering the boundary conditions at the surfaces of separation, find the amplitude of the reflected wave for $z < 0$. Show that this amplitude is zero if L equals a quarter of a guide wave length (in the material filling the section $0 < z < L$), and if $Z_1 = \sqrt{Z_0Z_2}$.

CHAPTER XII

SPHERICAL ELECTROMAGNETIC WAVES

Suppose we have an electric charge oscillating sinusoidally with the time. This charge will send out a spherical electromagnetic wave, radiating in all directions. Several physical problems are connected with such a wave. First, the phenomenon may be on a large scale, as in a radio antenna. Radiation from a vertical antenna, as a matter of fact, can be treated approximately by replacing the antenna by such an oscillating charge. But also on a smaller scale we can treat the radiation of short electromagnetic waves, or in other words light, from an atom that contains oscillating electrons. The electrons may have been set in motion by heat or bombardment, in which case we have the treatment of the emission of light from a luminous body; or they may be in forced motion under the action of another light wave, as in the case of the scattering of light. As a first step in the discussion of these problems, we consider the solution of Maxwell's equations in spherical coordinates. This solution, similar to the one we found in Chap. III for the static problem, will reduce in a simple case to the field of an oscillating dipole, which we require for the applications mentioned above. At the same time, it will give us the field of any oscillating multipole. Other forms of the same solution will lead to the field inside a spherical resonant cavity, or more generally in a cavity resonator bounded by surfaces r = constant, θ = constant, and φ = constant, in a system of spherical polar coordinates. We pass then to the solution of Maxwell's equations, later taking up their applications.

1. Maxwell's Equations in Spherical Coordinates.—In spherical polar coordinates, r, θ, φ, we can define a vector such as $\mathbf{E}$ in terms of its components E_r, E_θ, E_φ, along the directions in which these coordinates increase. Then, using the methods of vector operations in curvilinear coordinates, which are discussed in Appendix IV, Maxwell's equations for free space, in the absence of charge and current, become

$$\frac{1}{r \sin \theta}\left[\frac{\partial}{\partial \theta}(\sin \theta\, E_\varphi) - \frac{\partial E_\theta}{\partial \varphi}\right] + \mu_0 \frac{\partial H_r}{\partial t} = 0$$

$$\frac{1}{r \sin \theta}\frac{\partial E_r}{\partial \varphi} - \frac{1}{r}\frac{\partial (r E_\varphi)}{\partial r} + \mu_0 \frac{\partial H_\theta}{\partial t} = 0$$

$$\frac{1}{r}\frac{\partial(rE_\theta)}{\partial r} - \frac{1}{r}\frac{\partial E_r}{\partial \theta} + \mu_0 \frac{\partial H_\varphi}{\partial t} = 0$$
$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 H_r) + \frac{1}{r \sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, H_\theta) + \frac{1}{r\sin\theta}\frac{\partial H_\varphi}{\partial\varphi} = 0$$
$$\frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta\, H_\varphi) - \frac{\partial H_\theta}{\partial\varphi}\right] - \epsilon_0 \frac{\partial E_r}{\partial t} = 0$$
$$\frac{1}{r\sin\theta}\frac{\partial H_r}{\partial\varphi} - \frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r} - \epsilon_0\frac{\partial E_\theta}{\partial t} = 0$$
$$\frac{1}{r}\frac{\partial}{\partial r}(rH_\theta) - \frac{1}{r}\frac{\partial H_r}{\partial\theta} - \epsilon_0 \frac{\partial E_\varphi}{\partial t} = 0$$
$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 E_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, E_\theta) + \frac{1}{r\sin\theta}\frac{\partial E_\varphi}{\partial\varphi} = 0. \qquad (1.1)$$

The first three of these represent the equation curl $\mathbf{E} + \mu_0\, \partial\mathbf{H}/\partial t = 0$, the fourth is div $\mathbf{H} = 0$, the next three are curl $\mathbf{H} - \epsilon_0\, \partial\mathbf{E}/\partial t = 0$, and the last is div $\mathbf{E} = 0$, where as before the divergence equations contribute no new information not included in the others. The solutions of these equations represent, in general, waves propagated outward, or inward, along the radius. Thus the r direction can be considered the direction of propagation. As with the case of the wave guides considered in Chap. XI, there are two types of waves: transverse electric, or TE, in which the longitudinal component of the electric field, E_r, is zero; and transverse magnetic, TM, in which H_r is zero. We shall set up the solutions of both types, and consider their properties.

The solution of the wave-guide problem was simplified a great deal because we could assume from the outset that each component of the field varied exponentially along the z axis, the direction of propagation. It cannot be assumed in a corresponding way here that the components vary exponentially with r; as a matter of fact, their dependence on r is given by certain Bessel's functions, which approach an exponential form only for large values of r. Therefore we cannot at once carry through derivations like those of Eqs. (2.2), (2.3), and (2.4) of Chap. XI, expressing all the field components in the TE case algebraically in terms of H_z, and all the field components in the TM case in terms of E_z. Nevertheless an essentially equivalent discussion can be carried through, except that here the relations are not simply algebraic but involve differentiation.

Thus let us start with the TE case. We let H_r be a scalar function of r, θ, φ. It is slightly more convenient to operate, not with H_r, but with a quantity that we may denote u, defined by

$$u = rH_r. \qquad (1.2)$$

It will appear that the equation that must be satisfied by u is the wave equation, which in spherical polar coordinates becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2}-\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}=0. \quad (1.3)$$

We assume a sinusoidal dependence on time, varying according to the exponential $e^{j\omega t}$. The equation can then be solved by separation of variables, just as in Sec. 2, Chap. III, where we discussed the corresponding static problem. Letting u be of the form

$$u = R(r)\Theta(\theta)\Phi(\varphi)e^{j\omega t}, \quad (1.4)$$

we find that the functions satisfy the equations

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\left[\frac{\omega^2}{c^2}-\frac{l(l+1)}{r^2}\right]R=0, \quad (1.5)$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+\left[l(l+1)-\frac{m^2}{\sin^2\theta}\right]\Theta=0, \quad (1.6)$$

$$\frac{d^2\Phi}{d\varphi^2}+m^2\Phi=0, \quad (1.7)$$

similar to Eqs. (2.3), (2.4), and (2.1), of Chap. III. We shall discuss the solutions of these equations in the next section. Assuming that we have found u from them, we can now set up values for all the other field components.

It is not easy to proceed straightforwardly to a solution of Maxwell's equations. We find, however, the following equations for the field components in terms of u:

$$\begin{aligned}
H_r &= \frac{u}{r}\\
H_\theta &= \frac{1}{l(l+1)}\frac{1}{r}\frac{\partial}{\partial\theta}\frac{\partial}{\partial r}(ru)\\
H_\varphi &= \frac{1}{l(l+1)}\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}\frac{\partial}{\partial r}(ru)\\
E_r &= 0\\
E_\theta &= \frac{-j\omega\mu_0}{l(l+1)}\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}(ru)\\
E_\varphi &= \frac{j\omega\mu_0}{l(l+1)}\frac{1}{r}\frac{\partial}{\partial\theta}(ru).
\end{aligned} \quad (1.8)$$

By direct substitution in Maxwell's equations, (1.1), we can show that these equations are satisfied if u obeys the wave equation (1.3), or its equivalent equations (1.5), (1.6), (1.7). Thus we prove that the functions (1.8) are the solutions of the problem which we desire. We note that, instead of defining H_θ and H_φ directly in terms of u, we can instead set up the relations

$$\begin{aligned} H_\theta &= \frac{1}{j\omega\mu_0}\frac{1}{r}\frac{\partial}{\partial r}(rE_\varphi), \\ H_\varphi &= \frac{-1}{j\omega\mu_0}\frac{1}{r}\frac{\partial}{\partial r}(rE_\theta). \end{aligned} \tag{1.9}$$

We shall soon see that these equations form the analogy to the relation (2.4) of Chap. XI, giving the relation between the transverse components of $\mathbf{E}$ and $\mathbf{H}$ in a wave guide.

The relations for the transverse magnetic, TM, case are analogous to those just set up. In that case, we have a longitudinal component E_r of electric field, which we write as

$$E_r = \frac{v}{r}, \tag{1.10}$$

where v satisfies the same wave equation (1.3) that u satisfied. In terms of v, we can then write the other components as

$$\begin{aligned} E_\theta &= \frac{1}{l(l+1)}\frac{1}{r}\frac{\partial}{\partial\theta}\frac{\partial}{\partial r}(rv) \\ E_\varphi &= \frac{1}{l(l+1)}\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}\frac{\partial}{\partial r}(rv) \\ H_r &= 0 \\ H_\theta &= \frac{j\omega\epsilon_0}{l(l+1)}\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}(rv) \\ H_\varphi &= \frac{-j\omega\epsilon_0}{l(l+1)}\frac{1}{r}\frac{\partial}{\partial\theta}(rv) \end{aligned} \tag{1.11}$$

where we have the relations

$$\begin{aligned} E_\theta &= \frac{-1}{j\omega\epsilon_0}\frac{1}{r}\frac{\partial}{\partial r}(rH_\varphi) \\ E_\varphi &= \frac{1}{j\omega\epsilon_0}\frac{1}{r}\frac{\partial}{\partial r}(rH_\theta). \end{aligned} \tag{1.12}$$

We now have our complete set of equations for determining the fields, and shall consider their solutions in the next section.

2. Solutions of Maxwell's Equations in Spherical Coordinates.—To find the components of a field in spherical coordinates, we have

seen that we must first find a scalar u or v, satisfying the wave equation (1.3), which separates to give (1.5), (1.6), and (1.7). The dependence on angle, given in (1.6) and (1.7), is just as it is in the static case taken up in Chap. III, and as in that case we find that $\Theta = P_l^m(\cos\theta)$, $\Phi = \sin m\varphi$ or $\cos m\varphi$. The equation for the function R, however, is more complicated than in the static case, though it reduces to it for $\omega = 0$. If we make the substitution $R = f/\sqrt{r}$, where f is a function of r, we find that (1.5) is transformed to Bessel's equation for a Bessel function of order $l + \frac{1}{2}$. Thus we have $R = J_{l+1/2}(\omega r/c)/\sqrt{r}$ or $N_{l+1/2}(\omega r/c)/\sqrt{r}$.

It is customary to define spherical Bessel and Neumann functions $j_l(x)$ and $n_l(x)$ by the equations

$$j_l(x) = \sqrt{\frac{\pi}{2x}}\, J_{l+1/2}(x), \qquad n_l(x) = \sqrt{\frac{\pi}{2x}}\, N_{l+1/2}(x). \tag{2.1}$$

If l is an integer, which it will be in our applications, we can show, as is mentioned in Appendix VII, that j_l and n_l can be expressed in analytic form in terms of algebraic and trigonometric functions. For the first few functions we have

$$\begin{aligned} j_0(x) &= \frac{\sin x}{x}, & n_0(x) &= -\frac{\cos x}{x} \\ j_1(x) &= \frac{\sin x}{x^2} - \frac{\cos x}{x}, & n_1(x) &= -\frac{\sin x}{x} - \frac{\cos x}{x^2}. \end{aligned} \tag{2.2}$$

At large values of x, the terms in $1/x$ are the leading ones. In the general case, this leading term is given by

$$\begin{aligned} j_l(x) &\to \frac{1}{x}\cos\left(x - \frac{l+1}{2}\pi\right), \\ n_l(x) &\to \frac{1}{x}\sin\left(x - \frac{l+1}{2}\pi\right). \end{aligned} \tag{2.3}$$

In the opposite limit, as x tends to zero, we can expand in power series. In this case the leading term is

$$\begin{aligned} j_l(x) &\to \frac{x^l}{1\cdot 3\cdot 5\cdots(2l+1)} \\ n_l(x) &\to -\frac{1\cdot 1\cdot 3\cdot 5\cdots(2l-1)}{x^{l+1}}. \end{aligned} \tag{2.4}$$

In terms of these spherical Bessel functions, it is clear that our functions u or v can be written in the form

$$u \text{ or } v = \left[j_l\left(\frac{\omega r}{c}\right) \text{ or } n_l\left(\frac{\omega r}{c}\right)\right] P_l^m(\cos\theta)(\sin m\varphi \text{ or } \cos m\varphi). \tag{2.5}$$

The question of which functions to use in (2.5) depends on the boundary conditions to be satisfied, and this in turn depends on the problem that we are trying to solve. The simplest problem is the field within a cavity resonator bounded by a perfectly conducting spherical shell. In this case the field must be finite at the origin; therefore we can use only the functions j_l, not the Neumann functions n_l. Furthermore, the electric vector may have only an r component at the value of r giving the radius of the sphere. This means, for a TE mode, as we see from (1.8), that $j_l(\omega r/c)$ must be zero at the radius of the sphere, and for a TM mode, from (1.11), that the r derivative of $rj_l(\omega r/c)$ must be zero at this radius. These conditions easily determine the resonant frequencies of the various modes. We can almost as easily satisfy the boundary conditions for a resonator formed by the space between two concentric spheres. By making suitable combinations of $j_l(\omega r/c)$ and $n_l(\omega r/c)$, we see at once, from (2.2), that we can adjust the phase of the sinusoidal part of the function to any desired value. We then may choose the phase correctly to satisfy the boundary condition on one of the surfaces, and determine the frequency to satisfy the conditions on the other surface. These problems have an obvious analogy to the two-dimensional problems of the modes within a circular cylinder, and a coaxial line, discussed in the preceding chapter.

More interesting problems come when we take up the field in a region extending to infinity. In such cases, we are generally interested, not in standing waves, as in the solutions we have set up so far, but in traveling waves, traveling either inward or outward along r. Thus for instance in a problem of the emission of radiation from an atom or a radio antenna, we wish only waves traveling outward. It is easy to combine the Bessel and Neumann functions to secure traveling waves. The functions $j_l(x) \pm jn_l(x)$ [where we are to distinguish $j = \sqrt{-1}$ from $j_l(x)$] are called "spherical Hankel functions." From (2.3) we see that at large distances they approach the values

$$j_l(x) \pm jn_l(x) \rightarrow \frac{1}{x} e^{\pm j[x-(l+1)\pi/2]}. \tag{2.6}$$

Thus, if x is replaced by $\omega r/c$, we see that for the $+$ sign this represents the space part of a wave traveling in along r, and for the $-$ sign a wave traveling out, in each case with an amplitude that falls off inversely as the distance, so that the intensity, which is proportional to the square of the amplitude, falls off inversely as the square of the distance. For radiation from a source then, we choose the $-$ sign

in (2.6), and thus secure only an outgoing wave, without an incoming wave. We note that the Hankel functions, containing the Neumann functions, all have singularities at the origin. Thus an outgoing wave must either be propagated from a finite charge distribution, within which, since the charge density is different from zero, Maxwell's equations in the form (1.1) do not apply; in this case, a separate solution for the field within the charge distribution must be found, and it must be joined to the solution we have found; or the outgoing wave must be produced by a singularity of field, which can take the form of certain point charges, dipoles, and other multipoles. We shall take up examples of such singularities in a later section.

A very important form of problem is that of scattering of radiation by an object of some sort. If the object has spherical symmetry, as for instance if it is a small sphere of metal or of a dielectric, we can solve the problem of scattering by means of our spherical solutions of Maxwell's equations. For instance, let us consider scattering by a perfectly conducting sphere. Physically, we consider a plane wave incident on the sphere, and scattered spherical waves emerging from it. In this simple case, there is no mechanism for loss of energy, and all the energy scattered is made up from energy that strikes the sphere. We must make up the field by superposing solutions of the type we have considered. These solutions must satisfy boundary conditions of two sorts. First, at the surface of the sphere, the tangential components of electric field, and radial components of magnetic field, must be zero. To satisfy these, we must superpose Bessel and Neumann functions, in such a way as to make the function u equal to zero at the surface of the sphere for a TE mode, or the normal derivative of rv equal to zero for a TM mode. This can be done straightforwardly, and the result in general will not turn out to be a Hankel function, and hence will represent both an incoming and an outgoing spherical wave.

In addition to the boundary condition at the surface of the reflecting sphere, we must satisfy another condition, which requires more thought. We can easily state it physically: the disturbance must consist entirely of an incoming plane wave, plus outgoing spherical waves. To apply this condition, we must know how to describe a plane wave in terms of Bessel and Neumann functions. This is a rather involved problem, but in the similar case of a scalar wave equation, such as we should find for the scattering of sound, where we need only consider the pressure as a function of position, it becomes fairly simple. Considering only a scalar solution of the wave equation,

one can prove the following theorem:

$$e^{\pm jkr\cos\theta} = \sum_{l=0}^{\infty} (2l+1)(\pm j)^l P_l(\cos\theta) j_l(kr). \tag{2.7}$$

Here $P_l(\cos\theta)$ is the value which $P_l^m(\cos\theta)$ takes on when $m = 0$. We note that $r\cos\theta = z$, so that, if we set $k = \omega/c$, the function on the left of (2.7) represents the space-dependent function which, combined with a time exponential $e^{j\omega t}$, would represent a plane wave traveling along the $\mp z$ direction. We thus see how to expand such a plane wave in terms of spherical Bessel functions. It is natural that no Neumann functions come into this expansion, for they would introduce a singularity at the origin, which the plane wave does not have.

To satisfy our requirements, we must then build up a solution of our problem, consisting of the sum of the plane wave (2.7), and only outgoing spherical waves. That is, the term of the solution associated with a given l value, and with $m = 0$ (for, as we easily see, this problem will involve only terms for $m = 0$), will have the form

$$(2l+1)(\pm j)^l P_l(\cos\theta)\{j_l(kr) + \alpha_l[j_l(kr) - jn_l(kr)]\} \tag{2.8}$$

where α_l is a coefficient measuring the amplitude of the lth scattered-wave component. Thus we see that the ratio of the coefficients of $j_l(kr)$ and $n_l(kr)$ must be $(\alpha_l + 1)/(-j\alpha_l)$. On the other hand, this ratio has been determined quite independently from the condition that the solution satisfy the boundary condition at the surface of the reflecting sphere. Thus, by equating these two values of the ratio of coefficients, we can evaluate the α_l's, and find the intensity of the scattered radiation.

This procedure in principle is simple, but in practice a complete discussion of the problem is beyond the scope of this text. We recall that our field is really a vector field, so that instead of (2.7) we must use a corresponding vector formulation of a plane wave in terms of spherical waves. When we do this, and carry through the discussion, we find the following sort of result: The behavior of the scattered radiation depends a great deal on the ratio of the diameter of the scattering sphere to the wave length. For a sphere very small compared with the wave length, the sphere acts almost as if it were in a uniform field. It then acquires a dipole moment, as we have shown in Chap. III, Sec. 4. This dipole alone scatters radiation, and we shall give an elementary discussion of this case in a later section of the present chapter, showing the type of radiation emitted from a dipole. As the sphere gets larger, it scatters a more and more compli-

cated field, consisting now not only of dipole radiation, but of the fields of higher multipoles, which are discussed in Appendix VI. As the sphere becomes very large compared with a wave length, the outgoing wave takes on less and less of the character of a scattered wave, and approaches a reflected wave as predicted by geometrical optics for the simple optical problem of the reflection of a plane wave by a spherical mirror. In all these cases, we must consider not merely that part of the spherical wave which travels out in all directions, but in particular that part which travels in almost the same direction as the plane wave, after it has passed over the scattering obstacle. In this case, the scattered and incident waves interfere with each other, and it can be shown that they interfere destructively, so as to remove energy from the incident wave, just enough to supply the energy radiated out in the scattered wave. If the obstacle is small compared with the wave length, this interference fills a poorly defined region behind the obstacle, whose dimensions are determined by interference theory, but as the obstacle becomes large, the region of destructive interference becomes sharply defined, and becomes the ordinary shadow of geometrical optics. Around the edge of this shadow, however, there are diffraction fringes, alternations of intensity, resulting directly from the interference phenomenon.

All this interference and diffraction comes as part of a complete discussion of the problem of scattering of a plane wave by a sphere, and makes it clear why a complete treatment of the problem is beyond the scope of the present book. However, we shall handle interference and diffraction in an elementary way in a later chapter, bringing out many of the features that are present in the exact solution. The type of discussion sketched in the present section is important not only in electromagnetic theory, but in all other forms of wave theory as well. A simple example is the scattering of sound, as treated for example by P.M. Morse, *Vibration and Sound*, 2d ed. (McGraw-Hill Book Company, Inc., New York). More far-reaching is the application to quantum mechanics. The problem of scattering of an electron, neutron, or other colliding particle by an atom or nucleus can be handled to a first approximation, in wave mechanics, by a problem of scattering in a spherically symmetrical region where the index of refraction is a function of r. The mathematical framework of this problem is identical with that which we have sketched here, and a thorough understanding of the solution of the wave equation in spherical coordinates is an essential for the understanding of any type of scattering of waves by small obstacles.

3. The Field of an Oscillating Dipole.—In many ways the most important spherical wave physically is that produced by an oscillating electric dipole. This is the simplest TM wave. If we consider our solution (1.11), we see that the function v must depend on θ or φ, in order that any of the components of field may be different from zero. Thus we cannot have $l = 0$, since the spherical harmonic $P_0(\cos\theta)$ is a constant and does not depend on angles at all. The lowest value that l may have is 1. In this case, m can be 0 or 1. We consider the case $l = 1$, $m = 0$. This proves to be the simple dipole field. If we choose v to be

$$v = \cos\theta[j_1(kr) - jn_1(kr)],$$

or

$$v = \cos\theta\left[-\frac{1}{kr} + \frac{j}{(kr)^2}\right]e^{-jkr}, \tag{3.1}$$

where $k = \omega/c$, which is equivalent to it according to (2.2), we find for the field components

$$\begin{aligned} E_r &= ke^{j(\omega t-kr)}\cos\theta\left[-\frac{1}{(kr)^2} + \frac{j}{(kr)^3}\right] \\ E_\theta &= \frac{k}{2}e^{j(\omega t-kr)}\sin\theta\left[-\frac{j}{kr} - \frac{1}{(kr)^2} + \frac{j}{(kr)^3}\right] \\ H_\varphi &= \frac{k}{2}\sqrt{\frac{\epsilon_0}{\mu_0}}\,e^{j(\omega t-kr)}\sin\theta\left[-\frac{j}{kr} - \frac{1}{(kr)^2}\right]. \end{aligned} \tag{3.2}$$

The other field components are zero. These functions, of course, are to be multiplied by an arbitrary amplitude.

To understand the meaning of our solution, let us first consider the terms that are important at small distances, those in the highest inverse powers of r ($1/r^3$ for E_r and E_θ, $1/r^2$ for H_φ). We shall show that the terms in the electric force represent the field of an electric dipole at the origin and that the magnetic field is the field of the corresponding current element, derived from the time rate of change of the dipole moment, as found from the Biot-Savart law. We remember that a dipole of moment M has a potential

$$\psi = \frac{(M\cos\theta)}{4\pi\epsilon_0 r^2},$$

as we saw in Eq. (5.1), Chap. III. The field is then

$$\begin{aligned} E_r &= -\frac{\partial\psi}{\partial r} = \frac{2M}{4\pi\epsilon_0}e^{j\omega t}\frac{\cos\theta}{r^3}, \\ E_\theta &= -\frac{1}{r}\frac{\partial\psi}{\partial\theta} = \frac{M}{4\pi\epsilon_0}e^{j\omega t}\frac{\sin\theta}{r^3}. \end{aligned} \tag{3.3}$$

We observe that these are just a constant, $-jk^2M/2\pi\epsilon_0$, times the corresponding terms in (3.2). Similarly we note that a dipole whose moment was $Me^{j\omega t}$ would carry with it a current element equal to its time derivative, or to $j\omega Me^{j\omega t}$. This small current element would have a magnetic field H_φ, which would be given according to the Biot-Savart law, as described in Sec. 1, Chap. V, by

$$H_\varphi = \frac{j\omega M}{4\pi} e^{j\omega t} \frac{\sin\theta}{r^2}. \tag{3.4}$$

This is likewise $-jk^2M/2\pi\epsilon_0$ times the value from (3.2), at small distances. It is clear, then, that if we multiply (3.2) by this factor, we shall find a field that is a correct solution of Maxwell's equations at all distances, and yet that reduces to the field of a dipole of moment $Me^{j\omega t}$ at small distances, so that it must represent correctly the field of such a dipole. The field is

$$\begin{aligned}
E_r &= \frac{Mk^3}{4\pi\epsilon_0} e^{j(\omega t-kr)} \cos\theta \left[\frac{2j}{(kr)^2} + \frac{2}{(kr)^3}\right] \\
E_\theta &= \frac{Mk^3}{4\pi\epsilon_0} e^{j(\omega t-kr)} \sin\theta \left[-\frac{1}{kr} + \frac{j}{(kr)^2} + \frac{1}{(kr)^3}\right] \\
H_\varphi &= \frac{j\omega Mk^2}{4\pi} e^{j(\omega t-kr)} \sin\theta \left[\frac{j}{kr} + \frac{1}{(kr)^2}\right].
\end{aligned} \tag{3.5}$$

4. The Field of a Dipole at Large Distances.—We have seen that at sufficiently small distances the field of an oscillating dipole reduces to what we should compute by Coulomb's law and the Biot-Savart law. As kr becomes appreciable compared with unity (that is, as $2\pi r/\lambda$ becomes appreciable, or as r approaches $1/2\pi$ wave lengths) the remaining terms become important. This is an example of a general situation: at distances from an oscillating charge distribution that are small compared with wave length, we can use Coulomb's law and the Biot-Savart law. On the other hand, for distances large compared with a wave length, the situation is entirely reversed, and the first term in each expression in (3.5) becomes the important one. The leading terms in E_θ and H_φ vary as $1/r$, while the leading term of E_r goes as $1/r^2$, and therefore can be neglected at sufficiently large distances. The field, in other words, becomes transverse at large distances, with an amplitude inversely proportional to the distance. In this range it is ordinarily referred to as the "radiation field." We see that **E** and **H** become at right angles to each other, and check further that the ratio of E_θ to H_φ approaches the value $\sqrt{\mu_0/\epsilon_0}$, characteristic of plane waves in free space.

The intensity varies as $\sin^2 \theta$, so that the maximum radiation is at right angles to the axis of the dipole. To measure the intensity, we may compute the average value of Poynting's vector. We have

$$\begin{aligned} S_r &= \tfrac{1}{2} Re E_\theta \bar{H}_\varphi \\ &= \frac{1}{2} Re \frac{\omega M^2 k^5}{16\pi^2 \epsilon_0} \sin^2 \theta \left[\frac{1}{(kr)^2} - \frac{j}{(kr)^5} \right] \\ &= \frac{\mu_0 \sqrt{\epsilon_0 \mu_0}\, \omega^4 M^2 \sin^2 \theta}{32\pi^2 r^2}. \end{aligned} \tag{4.1}$$

We see that the intensity of radiation varies as $1/r^2$, the inverse-square law, as of course it must if the total flux outward through a sphere is independent of the size of the sphere. To find this total flux, we must integrate over all directions, by multiplying by the element of area $2\pi r^2 \sin \theta \, d\theta$ and integrating from 0 to π. We then have

$$\begin{aligned} \int S \, da &= \frac{\mu_0 \sqrt{\epsilon_0 \mu_0}\, \omega^4 M^2}{32\pi^2} 2\pi \int_0^\pi \sin^3 \theta \, d\theta \\ &= \frac{\mu_0 \sqrt{\epsilon_0 \mu_0}\, \omega^4 M^2}{12\pi}. \end{aligned} \tag{4.2}$$

This is a well-known formula for the radiation from a dipole. The two essential features are that the radiation is proportional to the square of the amplitude of the dipole, and to the fourth power of the frequency. It can be used for such diverse problems as finding the rate of radiation from an excited atom, or from a radio antenna, if the latter can be approximated, as it often can, as an oscillating dipole.

5. Scattering of Light.—In Sec. 3 we have given an indication of the theory of the scattering of light. We can take up easily a special and important case of scattering. Let a plane wave fall on the type of dipole that we have assumed in Chaps. IV and IX as being responsible for the dispersion and dielectric properties of matter. We assume the dipole to be small in dimensions compared with the wave length, so that the electric field acting on the dipole may be taken to be $E_0 e^{j\omega t}$, disregarding the variations with position. Let the dipole have an equation of motion

$$m \frac{d^2x}{dt^2} + mg \frac{dx}{dt} + m\omega_0^2 x = eE_0 e^{j\omega t}, \tag{5.1}$$

as in Eq. (1.2), Chap. IX. Then the dipole moment will be

$$Me^{j\omega t} = \frac{e^2}{m} \frac{E_0 e^{j\omega t}}{\omega_0^2 - \omega^2 + j\omega g}. \tag{5.2}$$

This is the oscillating dipole moment produced by the field.

The dipoles set into motion by the wave will emit light, which is scattered. The rate of emission by a single dipole is found by substituting the absolute value of the dipole moment in (5.2) into (4.2). Often the scattering is measured by the scattering cross section σ. This by definition is the area on which enough energy falls from the plane wave to equal the scattered intensity. Thus the rate of emission must equal σ times the Poynting vector of the plane wave, which is $\frac{1}{2}\sqrt{\epsilon_0/\mu_0}\,E_0^2$. Making the substitutions, we then find

$$\sigma = \frac{32\pi}{3} \frac{\omega^4}{[(\omega_0^2 - \omega^2)^2 + (\omega g)^2]} \left(\frac{e^2}{8\pi\epsilon_0 mc^2}\right)^2. \tag{5.3}$$

The quantity $(e^2/8\pi\epsilon_0 mc^2)$ is the quantity that we found, in Sec. 4, Chap. VIII, to be a classical radius R of the electron of charge e. Thus we find the interesting result that dimensionally the scattering cross section is determined by the square of this classical radius, or by the classical cross section of the electron, though with a numerical factor, and with a factor depending on frequency, which behaves quite differently in different parts of the spectrum. We shall now consider three important special cases of this scattering formula, holding for different frequencies:

a. The Rayleigh Scattering Formula.—This is what we have in the case where ω is small compared with ω_0. Since for ordinary atoms ω_0 is a frequency in the ultraviolet, we have this condition in the visible range of the spectrum. Then (5.3) becomes

$$\sigma = \frac{32\pi}{3} R^2 \left(\frac{\omega}{\omega_0}\right)^4. \tag{5.4}$$

The scattering is here proportional to ω^4, or to $1/\lambda^4$, where λ is the wave length. This is the Rayleigh scattering formula, developed to discuss the scattering of light by the sky. The proportionality to the inverse fourth power of the wave length means that the short blue and violet waves will be scattered by the air molecules much more than the long red ones, resulting in the blue color of the scattered light from the sky. The transmitted light thus has the blue removed and looks red, explaining the color near the sun at sunset.

b. The Thomson Scattering Formula.—In the other limiting case of X rays, when the frequency is large compared with ω_0, the scattering

becomes

$$\sigma = \frac{32\pi}{3} R^2, \tag{5.5}$$

the Thomson scattering formula. This formula gives a scattering independent of the wave length, and is very important in discussing X-ray scattering by substances.

c. Resonant Scattering.—If ω is nearly equal to ω_0, it is evident that the denominator can become very small, resulting in very large scattering. This phenomenon can be much more conspicuous than the other two cases. Thus a bulb filled with sodium vapor, which has a natural frequency in the visible region, illuminated with light of this color, will scatter so much light that it appears luminous. This phenomenon is called "resonance scattering." Here, as with the absorption lines discussed in Chap. IX, we have to correlate our classical theory of electric dipoles with the quantum theory. Actual atoms scatter, as they absorb, at frequencies connected with quantum transitions between energy levels. However, the same description of the dipoles that is correct for describing absorption and dispersion is also applicable to scattering.

One observation regarding scattering is that, if the incident light is plane polarized, the dipoles will all vibrate along the direction of its electric vector. Thus there will be no intensity in the scattered light along this direction. The scattered light will have a maximum intensity at right angles, and it will be plane polarized. It was by experiments based on these facts that the polarization of X rays was first found.

6. Coherence and Incoherence of Light.—In the preceding paragraphs, we have calculated the scattering from a single dipole or a single atom, when a plane wave falls on it. It is found in practice that, in a gas containing N molecules per unit volume, the total intensity scattered by unit volume will be just N times that scattered by a single molecule. On the other hand, if the molecules are regularly arranged in a crystalline solid, there is practically no scattering. The effect rather is the dispersion and absorption discussed in Chap. IX. We must give closer consideration to the question of why this distinction exists, and why in particular the intensity of the scattered radiation from N molecules of a gas is N times that from a single molecule. Since the Maxwell equations are linear, the field vectors **E** and **H** satisfy the superposition principle, so that we should expect the total amplitude to be the sum of the amplitudes in the various waves, in which case the total intensity, being the square of the ampli-

tude, would certainly not be the sum of the separate intensities. The key to this situation is found in the relations between the phases of the various waves that we are adding: if they all have fixed phases relative to each other, they are said to be "coherent," and the amplitudes add; whereas if they are in phases having random relations to each other they are "incoherent," and the intensities add.

To be more precise, let us consider the sum of a number of sinusoidal waves, all of the same frequency, but of different amplitude and phase:

$$\sum_k A_k \cos (\omega t - \alpha_k) = \left(\sum_k A_k \cos \alpha_k\right) \cos \omega t + \left(\sum_k A_k \sin \alpha_k\right) \sin \omega t.$$

If all the phases should be the same, say $\alpha_k = 0$, then the amplitudes of the cosine and sine terms will be $\sum_k A_k$ and 0, respectively, so that the amplitudes add, and the intensity is proportional to $\left(\sum_k A_k\right)^2$, or, if for instance there are N terms of equal amplitude, proportional to N^2 times the intensity of a single wave. On the other hand, the α's may be completely independent of each other, meaning that each α is equally likely to have any value between 0 and 2π, independent of the others. Then we can see that $\sum_k A_k \cos \alpha_k$ will be far less than $\sum_k A_k$, since we shall have just about as many terms with positive values of $\cos \alpha_k$ as with negative, and the terms will just about cancel.

The cancellation will not be complete, however, as we see if we compute the squares of the summations, which we must add to get the intensity. The square of the first summation, for instance, is

$$\left(\sum_k A_k \cos \alpha_k\right)^2 = \sum_k A_k^2 \cos^2 \alpha_k + \sum_{k \neq l} A_k A_l \cos \alpha_k \cos \alpha_l.$$

We must find the average of this, taking the α's as independent. That is, we must perform the operation of integrating each α from 0 to 2π, and dividing by 2π. When we do this, the terms $\cos^2 \alpha_k$ average to $\frac{1}{2}$, while the products of two independent α's average to zero, leaving $\frac{1}{2} \sum_k A_k^2$. The other summation gives an equal term, so that we find that the mean-square amplitude, or mean intensity, averaged over phases, is the sum of the individual intensities. This is the state of complete incoherence, in which for N waves the intensity

is N times the intensity of a single wave, rather than N^2 as for the coherent case. The cancellation of waves, then, while not complete, is more and more perfect as N increases, for N becomes a smaller and smaller fraction of N^2 as N increases.

We can now apply the idea of coherence to the scattering of light from a gas. The phase of the wave at a point P, scattered by an atom at a, in Fig. 27, depends on the total path the light has traveled from the source to a, and from a to P. Since the molecules of a gas have no fixed positions with respect to each other, these paths are in a random relation to each other, the phases are incoherent, and we are justified in adding intensities. Such a procedure would not be allowed for example in discussing the scattering of X rays by crystals,

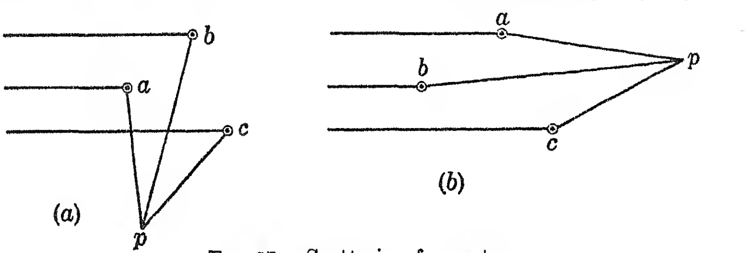


FIG. 27.—Scattering from atoms.

where the various atoms are in fixed lattice positions. Indeed, here we do get interference, and it is just by studying the interference patterns so obtained that we obtain our information about the lattice structure of crystals. Neither would the procedure be allowed in discussing the scattering from a gas in the same direction as the incident radiation, as in (b), Fig. 27. For then the paths of the beams scattered from the various atoms are approximately equal, the waves are in phase, and they produce a resultant field at P proportional to the amplitude, rather than the intensity, of the incident wave. This scattered field can be shown to interfere with the incident wave in such a way that the resultant produces the refracted wave. The close relation of our scattering formulas to the formulas for the index of refraction, therefore, becomes clear, and it is evident that our two problems of refraction and scattering, though we have treated them separately, are really parts of the same subject. Scattering straight ahead produces refraction, and does not depend on exact placing of the molecules. Scattering to the sides, on the other hand, does not occur unless the molecules have a random arrangement; then the intensity, not the amplitude, is proportional to the number of molecules.

We have just been considering the coherence of scattered light.

There is another aspect of coherence, related to this, which can be discussed in a similar way. This is the incoherence between different independent sources of light, even when they have the same frequency. It is based on the relation between coherence and the spectrum. The amplitude of a wave of light, as a function of time, is never exactly sinusoidal, but is really a much more complicated function. It is often desirable, however, to resolve such a function into a spectrum; that is, write it as a sum of sinusoidal waves of different frequency. This can be conveniently done by Fourier series, as is discussed in Appendix III.

To do this, we take a Fourier series with an extremely long period T, so long that all the phenomena we are interested in take place in a time short compared with T, so that we are not bothered by the periodicity of the series. Then, if our function is $f(t)$, we have

$$f(t) = \sum_n (A_n \cos \omega_n t + B_n \sin \omega_n t),$$

where

$$A_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t) \cos \omega_n t \, dt, \qquad B_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t) \sin \omega_n t \, dt,$$

$$\omega_n = \frac{2\pi n}{T}.$$

This gives an analysis into an infinite number of sine waves, with frequencies spaced very close together (on account of the very small size of $2\pi/T$). No actual, physical wave is then perfectly sinusoidal, in the sense of having but one term in this expansion with an amplitude different from zero. We shall show in a problem that even a perfectly sinusoidal wave that persists for only a finite length of time will have appreciable amplitudes for all those frequencies within a range $\Delta\omega$, equal in order of magnitude to the reciprocal of the time during which the wave persists, so that a sine wave of long lifetime will correspond to a sharp line in the spectrum, while a rapidly interrupted wave will give a broad line. This is observed experimentally in the fact that increasing the pressure of a gas, thereby making collisions more frequent and interrupting the radiating of the atoms, broadens the spectral lines.

The intensity is proportional to $f^2(t)$, or to the square of the summation over frequencies. Just as before, this square consists of terms like $A_n^2 \cos^2 \omega_n t$, and cross terms like $A_n A_m \cos \omega_n t \cos \omega_m t$. Instantaneously none of these terms are necessarily zero. But if we average over time, the terms of the first sort average to $A_n^2/2$, while those of the second sort average to zero. The final result, then is that the time

average intensity is the sum of the intensities of the various frequencies: $\overline{f^2(t)} = \frac{1}{2} \sum_n (A_n^2 + B_n^2)$. We are justified in considering the terms connected with a given n to be the intensity of light of that particular frequency in the spectrum, so that we have the theoretical method of determining the spectral analysis of any disturbance. And we see that the following statement is true: on a time average, sinusoidal waves of different frequencies are always incoherent, and never interfere.

It is known experimentally that light from two different sources never interferes; to get interference we must take light from a single source, split it into two beams, and allow these beams to recombine. If we regarded the sources as being monochromatic, it would be hard to see why this should be, for the amplitudes of two waves of the same frequency should add, rather than the intensities, and this is the essence of interference. But when we observe that each source really is represented by a Fourier series, the situation becomes plain. For two sources are always so different that their Fourier series will be entirely different. If we analyze both of them, the phase of the radiation of frequency ω_n from one will be entirely independent of the phase of the corresponding frequency from the other. Thus if we add, square, and average over this random relation between the phases of the two sources, the cross terms will cancel, and the intensities add. The randomness comes in this case, not in adding a great many terms of the same frequency, but in combining the terms of different frequencies, which are related in entirely independent ways in the two sources.

Problems

1. Discuss the weakening of sunlight because of scattering, as the light passes through the atmosphere. Assume that the molecules of the atmosphere have a natural frequency at 1,800 A (where absorption is observed). Let each molecule contain an electron of this frequency. Assume that the number of molecules is such as to give the normal barometric pressure. Find the fractional weakening of a beam due to scattering in passing through a sheet of thickness ds, and from this set up the differential equation for intensity as a function of the distance. Solve for the ratio of intensity to the intensity before striking the atmosphere, for the sun shining straight down, and for it shining at an angle of incidence of 60°. Constants: $e = 1.60 \times 10^{-19}$ coulomb, $m = 9.1 \times 10^{-31}$ kg, number of molecules in 1 gm-mol $= 6.03 \times 10^{23}$.

2. A vibrating dipole radiates energy, and therefore its own energy decreases. Noting that the rate of radiation is proportional to the energy, set up the differential equation for the energy of the dipole as a function of the time. Find how long it takes the dipole to lose half its energy. Work out numerical values for the sort of dipole considered in Prob. 1.

3. Using the results of Prob. 2, find the equivalent damping term that would make the dipole lose energy at the same rate as the radiation. This damping is called the "radiation resistance."

4. Suppose we have an alternating current of maximum value I, in a vertical antenna of length L. Treating this as a dipole, find the total radiation. Show that the equivalent resistance necessary to produce the same power loss (the radiation resistance) is

$$R = \frac{2\pi}{3}\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{L^2}{\lambda^2} \qquad \text{ohms.}$$

5. Show that the field of an oscillating dipole of moment $Me^{j\omega t}$, pointing along the axis of the spherical coordinates, can be derived from the following scalar and vector potentials:

$$\psi = -\frac{M}{4\pi\epsilon_0}\frac{d}{dr}\left[\frac{e^{j\omega(t-r/c)}}{r}\right]\cos\theta,$$

$$A_r = \frac{j\omega\mu_0 M e^{j\omega(t-r/c)}}{4\pi r}\cos\theta$$

$$A_\theta = -\frac{j\omega\mu_0 M e^{j\omega(t-r/c)}}{4\pi r}\sin\theta.$$

$$A_\varphi = 0.$$

6. The so-called "Hertz vector" arising from a dipole of moment $\mathbf{M}e^{j\omega t}$, at a distance r from the dipole, is equal to

$$\mathbf{\Pi} = \frac{\mathbf{M}e^{j\omega(t-r/c)}}{4\pi\epsilon_0 r},$$

and points in the direction of the dipole moment. Show that in terms of the Hertz vector the potentials can be found from the expressions

$$\mathbf{A} = \epsilon_0\mu_0\frac{\partial\mathbf{\Pi}}{\partial t}, \qquad \psi = -\operatorname{div}\mathbf{\Pi}.$$

7. Show from the Hertz vector that the field at large distances from a dipole is given by

$$\mathbf{E} = \sqrt{\frac{\mu_0}{\epsilon_0}}\frac{1}{4\pi c r^3}\left\{\mathbf{r}\times\left[\mathbf{r}\times\frac{d^2}{dt^2}\mathbf{M}e^{j\omega(t-r/c)}\right]\right\},$$

$$\mathbf{H} = -\frac{1}{4\pi c r^2}\left[\mathbf{r}\times\frac{d^2}{dt^2}\mathbf{M}e^{j\omega(t-r/c)}\right].$$

8. Find the spectrum of a disturbance that is zero up to $t = 0$, is sinusoidal until $t = T_0$, then is zero permanently. (*Hint:* Make the period T of the Fourier series indefinitely large compared with T_0.)

9. Find the spectrum of a disturbance that starts at $t = 0$, and is a sinusoidal damped wave after that. Show that the curve for intensity as a function of frequency has the same form as a resonance curve, in general, and that its breadth is connected with the damping constant in the same way. This illustrates an important principle: the emission and absorption spectra of the same substance are essentially equivalent. The resonance curve represents the absorption curve, because of the relation of forced oscillators and dispersion, whereas the damped wave is the emission. (*Hint:* Make the period T indefinitely large compared with the time taken for the oscillation to $1/e$th of its value.)

CHAPTER XIII

HUYGENS' PRINCIPLE AND GREEN'S THEOREM

Huygens' principle provides a well-known elementary method for treating the propagation of waves, and in this chapter we shall consider its mathematical background, showing its close connection with Green's theorem. The method is this: From each point of a given wave front, at $t = 0$, we assume that spherical wavelets originate. At time t, each wavelet will have a radius ct, and the envelope of these wavelets will form a new surface, which according to Huygens is simply the resulting wave front at this later time t. Thus, if the original wave front was a plane, it is easy to see that the final one will be a plane distant by the amount ct, while if it is a sphere, the final wave front will be a concentric sphere whose radius is larger by ct. In either case this construction gives us the correct answer, agreeing with the more usual methods of computation. The one difficulty is that our construction would give a wave traveling backward, as well as one traveling forward; the solution of this difficulty appears when we use the methods of the present chapter.

We may look at our process in a slightly different way, not used by Huygens, but developed early in the nineteenth century when the interference of light was being worked out by Young and Fresnel. Suppose that, instead of taking the envelope of all the spherical wavelets, we consider that each of these wavelets has a certain amplitude, consisting of a sinusoidal vibration. We then add these vibrations, just as if the wavelets were being sent out by interfering sources of light, and the resulting amplitude is taken to be that in the actual wave. This process can be shown to lead to essentially the same result, and it is this which can be justified theoretically. As a further generalization, it is not necessary to take the original surface to be a wave front; it can be any closed surface, so long as we allow the scattered wavelets to have the suitable phase and amplitude.

Our final result, then, is this: The disturbance at a point P of a wave field may be obtained by taking an arbitrary closed surface, and performing an integration over this surface. The contribution of a small element of area da of this surface equals the amplitude at P of a spherical wave starting from da at such a time that it reaches P at

time t. The simplest form of spherical wave is one we did not mention in Chap. XII. It is a scalar solution of the wave equation, independent of angles, and depending only on r. The wave equation (1.3) of Chap. XII, in a case where u depends only on r, may be written

$$\frac{\partial^2(ru)}{\partial r^2} - \frac{1}{c^2}\frac{\partial^2(ru)}{\partial t^2} = 0,$$

which has solutions $u = \dfrac{f(t \pm r/c)}{r}$, where f is an arbitrary function. If we have a simple spherical wave of this type, we shall choose the $-$ sign, so as to represent a wave traveling out along the radius with velocity c, and of amplitude varying inversely as the distance. Now the contribution of an element da must surely be proportional to the disturbance at da, which we may call f (a function of time and position), and to da. Hence we have something like $\displaystyle\int \frac{f(t - r/c)}{r}\,da$ for the final result. We are thus led to a formula of this sort:

$$f\text{ (at a point } P) = \text{constant} \times \int \frac{f(t - r/c)}{r}\,da,$$

where the surface integral is over a surface surrounding P. This suggests the solution of Laplace's equation by Green's method, where we obtained the value of a function φ at an interior point of a region where $\nabla^2\varphi$ was zero as a surface integral over the boundary. As a matter of fact, an analogue to Green's theorem is the correct statement of Huygens' principle, and replaces the formula that we derived intuitively above, and that is not just correct. We shall now proceed to set up this theorem, first introducing the retarded potentials, which are closely related to the problem.

1. The Retarded Potentials.—In Chap. VII, we introduced scalar and vector potentials, φ and $\mathbf{A}$, giving the electric and magnetic fields by the relations

$$\mathbf{E} = -\operatorname{grad}\varphi - \frac{\partial \mathbf{A}}{\partial t}$$
$$\mathbf{B} = \operatorname{curl}\mathbf{A}.$$

For these potentials, in empty space, we found the equations

$$\nabla^2\varphi - \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = -\frac{\rho}{\epsilon_0}$$
$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu_0\mathbf{J}, \qquad (1.1)$$

or d'Alembert's equation. We first ask how to get a solution of d'Alembert's equation analogous to the simple solution

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{\rho}{r}\,dv = -\frac{1}{4\pi}\int \frac{\nabla^2\varphi}{r}\,dv \tag{1.2}$$

of Poisson's equation. We shall not carry through the proof of the solution, for that is rather complicated. But the essence of the solution of Poisson's equation is that we divide up all space into volume elements dv, and that $\rho\,dv/4\pi\epsilon_0 r$ is the potential of the point charge $\rho\,dv$ at a distance r. This potential, of course, is a solution of Laplace's equation, as is $1/r$, at all points except for $r = 0$, where the charge is located.

In a similar way, to solve d'Alembert's equation, we divide up our charge into small elements, and write the potential as the sum of the separate potentials of these small charges. The separate potentials must now be, except at $r = 0$, solutions of the wave equation. This means that, since any change of the charge will be propagated outward with the velocity c, the potential at a given point of space resulting from a particular charge cannot be derived from the instantaneous value of the charge, but must be determined, instead, by what the charge was doing at a previous instant, earlier by the time r/c required for the light to travel out from the charge to the point we are interested in. In other words, if $\rho(x,y,z,t)$ is the charge density at x, y, z at the time t, and r is the distance from x, y, z to x', y', z' where we are finding the field, we shall expect the potential of the charge in dv to be

$$\frac{\rho(x,y,z,t - r/c)\,dv}{r}$$

and for the whole potential we shall have

$$\begin{aligned}\varphi &= \frac{1}{4\pi\epsilon_0}\int \frac{\rho(x,y,z,t - r/c)\,dv}{r} \\ &= -\frac{1}{4\pi}\int \frac{[\nabla^2\varphi - (1/c^2)(\partial^2\varphi/\partial t^2)]_{t-r/c}}{r}\,dv.\end{aligned} \tag{1.3}$$

This solution is, as a matter of fact, correct. We have already seen that $\dfrac{f(t - r/c)}{r}$ is a solution of the wave equation, where f is any function, so that the integrand actually satisfies the wave equation, as in the earlier case $1/r$ satisfied Laplace's equation. The potential φ determined by this equation is called a "retarded poten-

tial," since any change in the charge is not instantaneously observable in the potential at a distant point, but its effect is retarded because of the finite velocity of light. The solution for each component of the vector potential is determined in an analogous manner.

2. Mathematical Formulation of Huygens' Principle.—In discussing the application of Green's theorem to the solution of Poisson's equation in a finite region of space, we proved in Sec. 6, Chap. III, that

$$\varphi = -\frac{1}{4\pi}\int \frac{\nabla^2\varphi}{r}\,dv - \frac{1}{4\pi}\int \left(\varphi \text{ grad } \frac{1}{r}\cdot \mathbf{n} - \frac{1}{r}\text{ grad } \varphi \cdot \mathbf{n}\right) da, \quad (2.1)$$

the result of the last paragraph being the special case where the region of integration is infinite and the surface integral drops out. We now wish to find an analogous theorem for use with d'Alembert's equation. Here again we shall not give a real derivation, for this is complicated, but shall merely describe the formula that results, and show that it is plausible. We have already discussed the volume integral. In the surface integral, the first term gave the potential of a double layer of strength proportional to $\varphi/4\pi$, the second the potential of a surface charge of magnitude proportional to

$$\left(\frac{1}{4\pi}\right)\text{grad } \varphi \cdot \mathbf{n}, \qquad \text{or} \qquad \left(\frac{1}{4\pi}\right)\left(\frac{\partial\varphi}{\partial n}\right).$$

Each of the terms, φ grad $(1/r)\cdot \mathbf{n}$, or $\varphi[\partial(1/r)/\partial n]$ and $(1/r)(\partial\varphi/\partial n)$ is a solution of Laplace's equation since it represents the potential of certain charges.

In our case of the wave equation, the formula has two corresponding terms: one giving the potential of a double layer, the other of a surface charge. But now the charges change with time, so that we must use solutions of the wave equation in the integral. We have already seen that the solution of the wave equation corresponding to $1/r$ is $f(t - r/c)/r$; hence we expect the second term to be replaced by $-(1/r)(\partial\varphi/\partial n)_{t-r/c}$, where this means that the partial derivative, which is now a function of time as well as of position on the surface, is to be computed, not at t, but at $t - (r/c)$. Similarly corresponding to $\partial(1/r)/\partial n$, the difference of the potentials of two equal and opposite point charges at neighboring points of space, we have

$$\frac{\partial}{\partial n}\left[\frac{f(t - r/c)}{r}\right].$$

Remembering that in differentiating with respect to n we must regard r as a variable each time it occurs, this is

$$f\left(t - \frac{r}{c}\right)\frac{\partial(1/r)}{\partial n} + \frac{1}{r}\frac{\partial}{\partial n}\left[f\left(t - \frac{r}{c}\right)\right]$$
$$= -\frac{\cos(n,r)}{r}\left[\frac{f(t - r/c)}{r} + \frac{1}{c}\frac{\partial f(t - r/c)}{\partial t}\right]$$

where in the last term we have used the relation

$$\begin{aligned}\frac{\partial f(t - r/c)}{\partial n} &= \frac{df(t - r/c)}{d(t - r/c)}\frac{\partial(t - r/c)}{\partial n} \\ &= \frac{\partial f(t - r/c)}{\partial t}\left(-\frac{1}{c}\frac{\partial r}{\partial n}\right) \\ &= -\frac{1}{c}\cos(n,r)\frac{\partial f(t - r/c)}{\partial t}.\end{aligned}$$

We should therefore expect to have

$$\begin{aligned}\varphi = -\frac{1}{4\pi}\int \frac{[\nabla^2\varphi - (1/c^2)(\partial^2\varphi/\partial t^2)]_{t-r/c}}{r}\,dv \\ + \frac{1}{4\pi}\int \frac{1}{r}\left\{\left[\frac{1}{c}\left(\frac{\partial\varphi}{\partial t}\right)_{t-r/c} + \frac{\varphi(t - r/c)}{r}\right]\cos(n,r)\right. \\ \left. + \left(\frac{\partial\varphi}{\partial n}\right)_{t-r/c}\right\}da. \quad (2.2)\end{aligned}$$

This, as a matter of fact, is the correct formula. The first term represents the potential due to all the charge within the volume; if there are no sources of radiation within this volume, the volume integral is then zero, and that is the usual case with optical applications. The surface integral represents the remaining potential as arising from a distribution of charge and double distribution about the surface, each surface element sending out a wavelet that on closer examination proves to be the Huygens' wavelet we are interested in. Thus, starting from Green's theorem and d'Alembert's equation, we have arrived at a mathematical formulation of Huygens' theorem.

To give a suggestion of the rigorous proof of this formula, we could proceed as follows: First, we notice that φ defined by this integral satisfies the wave equation; for since each term of the integrand separately is a solution, the sum must also be. Now it follows from this, although we have not proved it, that if the solution reduces to the correct boundary values at all points of the boundary, the solution must be the correct one, the reason being essentially that the boundary values determine a solution uniquely, so that, if we have one solution of the equation with the right boundary values, it must be the only correct solution. We must then show that the φ defined by the

integral actually has the correct boundary values. This could be done by a more careful treatment, and we should then have a demonstration of the formula. The more conventional proof, however, is a fairly direct though complicated application of Green's theorem.

We shall now take our general formula (2.2), and apply it to the cases we meet in optics, showing that it reduces to something like the formula that we derived earlier intuitively. We suppose that light is emitted by a point source, and that the value of some quantity connected with it, and satisfying the wave equation (one of the components of the fields or potentials—they all satisfy the same relations) has the form $\frac{Ae^{j\omega(t-r_1/c)}}{r_1}$, where r_1 is the distance from the source to the point where we wish to find the disturbance. Then we wish to get the disturbance at P, not by direct calculation, but by using Huygens' principle. Suppose we take a closed surface. This surface can surround either the source, or the point P where we wish the disturbance. In any case, we define $\mathbf{n}$ as the normal pointing out of the part of the space in which P is located. At a point of the surface, $\varphi = \frac{Ae^{j\omega(t-r_1/c)}}{r_1}$, where r_1 is the distance from the source to the point on the surface.

We then have, if r is the distance from P to a point on the surface,

$$\varphi\left(t - \frac{r}{c}\right) = \frac{Ae^{j\omega[t-(r+r_1)/c]}}{r_1}$$

$$\frac{\partial\varphi(t - r/c)}{\partial t} = \frac{j\omega Ae^{j\omega[t-(r+r_1)/c]}}{r_1}$$

$$\frac{\partial\varphi(t - r/c)}{\partial n} = -A\cos(n,r_1)\left(\frac{1}{r_1} + \frac{j\omega}{c}\right)\frac{e^{j\omega[t-(r+r_1)/c]}}{r_1}.$$

Thus finally, substituting in (2.2) and rearranging terms,

$$\varphi = \frac{1}{4\pi}\int \frac{A}{rr_1}e^{j\omega[t-(r+r_1)/c]}\left[\left(\frac{1}{r} + \frac{j\omega}{c}\right)\cos(n,r) - \left(\frac{1}{r_1} + \frac{j\omega}{c}\right)\cos(n,r_1)\right]da. \quad (2.3)$$

In this formula, as in Chap. XII, we have two sorts of terms, some significant at small, others at large values of r and r_1. We easily see that, if r and r_1 are large compared with a wave length, as is always the case in optics, the only terms we need retain are those in $j\omega/c$. Hence to this approximation

$$\varphi = \int \frac{jA}{2\lambda r_1 r}e^{j\omega[t-(r+r_1)/c]}[\cos(n,r) - \cos(n,r_1)]\,da. \quad (2.4)$$

This final form suggests our earlier, intuitive formulation of Huygens' principle. The incident amplitude at da is $\dfrac{Ae^{j\omega(t-r_1/c)}}{r_1}\cdot$ Now we set up, starting from da, a wavelet whose amplitude is this value, retarded by the amount r/c, divided by r, and multiplied by the factor $(j/2\lambda)[\cos(n,r) - \cos(n,r_1)]\,da$. This is just what we should expect, except for the last factor. The term j introduces a change of phase of 90°, not present in Huygens' form of the principle, but necessary. The term $\cos(n,r) - \cos(n,r_1)$ makes the wavelets have an amplitude that depends on angle. When r and r_1 are in opposite directions, which is the case when the surface is between the source and P, the factor approaches 2; whereas, when r and r_1 are parallel, and the surface is beyond P, it becomes zero. This means that the wavelets do not travel backward, thus removing the difficulty noticed earlier in Huygens' method. The wavelets have an amplitude depending on their wave length, decreasing for the longer wave lengths.

3. Integration for a Spherical Surface by Fresnel's Zones.—Let us now carry out our integration, and verify Huygens' method, in a simple case. We take the surface to be a sphere, surrounding the source, and therefore a wave front. We note that n is the inner normal of the sphere. Thus r_1 is constant all over the sphere, and $\cos(n,r_1) = -1$ at all points, so that the formula simplifies to

$$\varphi = \frac{jAe^{j\omega(t-r_1/c)}}{2\lambda r_1}\int \frac{e^{-jkr}}{r}[\cos(n,r) + 1]\,da,$$

where $k = \omega/c = 2\pi/\lambda$. Now suppose we introduce, as a coordinate on the sphere, the distance r from the point P; that is, we cut the sphere with spheres concentric with P, laying off zones between them, as in Fig. 28. We can easily get the area between r and $r + dr$, and hence

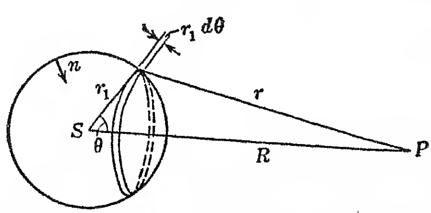


FIG. 28.—Construction for Fresnel's zones on a sphere surrounding the source.

the element of area. Take as an axis the line joining the source and the point P, and consider a zone making an angle between θ and $\theta + d\theta$ with the axis. The area of the zone is $2\pi r_1^2 \sin\theta\,d\theta$. But now

by the law of cosines, if R is the distance from the source to P,

$$r^2 = R^2 + r_1^2 - 2Rr_1 \cos \theta,$$

and differentiating, $2r\,dr = 2Rr_1 \sin \theta\, d\theta$. Hence for the area of the zone we have $\frac{2\pi r r_1}{R}\,dr$. Introducing this, we have

$$\varphi = \frac{j\pi A e^{j\omega(t-r_1/c)}}{\lambda R} \int_{r_{\min}}^{r_{\max}} e^{-jkr}[\cos (n,r) + 1]\, dr,$$

where $r_{\min} = R - r_1$, $r_{\max} = R + r_1$.

To carry out this integration, we use a device called "Fresnel's zones," giving us an approximate value in a very elementary way. Beginning with $r_{\min}$, we take a set of zones such that the outer edge of each corresponds to a value of r just half a wave length greater than the inner edge. The contributions of successive zones will almost exactly cancel. The integral, then, consists of a sum of terms, say $s_1 - s_2 + \cdots + s_n$, where the magnitudes of $s_1, s_2, \cdots$, vary only slightly from one to the next. Now it is true in general that in such a series the sum is approximately half the sum of the first and last terms. We can see this as follows: We group the terms

$$\frac{s_1}{2} + \left(\frac{s_1}{2} - s_2 + \frac{s_3}{2}\right) + \cdots + \left(\frac{s_{n-2}}{2} - s_{n-1} + \frac{s_n}{2}\right) + \frac{s_n}{2}.$$

Now, because of the slow variation of magnitude, we have very nearly $s_k = \frac{s_{k-1} + s_{k+1}}{2}$. If this were so, however, each of the parentheses would vanish, leaving only $\frac{s_1 + s_n}{2}$.

In our case, the contribution of the first zone is to be considered, but that of the last zone is practically zero, because of the factor $\cos (n,r) + 1$, so that the result is half the first zone. Now, in the first zone, $\cos (n,r) + 1$ is so nearly equal to 2 that we can take it outside the integral, obtaining

$$\begin{aligned}
\varphi &= \frac{\pi j}{\lambda} \frac{A e^{j\omega(t-r_1/c)}}{R} \int_{R-r_1}^{R-r_1+\lambda/2} e^{-jkr}\, dr \\
&= -\frac{A e^{j\omega(t-r_1/c)}}{2R} \left(e^{-jkr}\right) \Big|_{R-r_1}^{R-r_1+\lambda/2} \\
&= \frac{A e^{j\omega(t-r_1/c)}}{R} e^{-jk(R-r_1)} \\
&= \frac{A e^{j\omega(t-R/c)}}{R},
\end{aligned} \tag{3.1}$$

the correct value.

In the derivations of this chapter we have traveled in a very roundabout way to reach a very obvious result. We naturally ask, what is Huygens' principle good for, aside from a mathematical exercise? The answer is found in problems of diffraction. There one has certain opaque screens, with holes in them, and a light wave falling on them. If the light comes from a point source, geometrical optics would tell us that the shadow of the screen would have perfectly sharp edges. But actually this is not true; there are light and dark fringes around the edge of the shadow. If the shadow is observed at a greater and greater distance, these fringes get proportionally larger and larger, until they entirely fill the image of the hole. Finally at great distances the fringes grow in size until the resulting pattern has no resemblance at all to the geometrical image. There are then two general sorts of diffraction: first, that in which the pattern is like the geometrical image, but with diffuse edges, and which is called "Fresnel diffraction"; secondly, that in which the pattern is so extended that it has no resemblance to the geometrical image, and which is called "Fraunhofer diffraction." Both types of diffraction, as well as the intermediate cases, can be treated by using Huygens' principle.

4. Huygens' Principle for Diffraction Problems.—Suppose that light from a point source falls on a screen containing apertures, and that we wish the amplitude at points behind the screen. Then we surround the point P, where we wish the field, by a surface consisting of the screen, and of a large surface, perhaps hemispherical, extending out beyond P, and enclosing a volume completely. We apply Huygens' principle to the surface. In doing so we assume (1) that the amplitude of the incident wave, at points on the apertures, is the same that it would be if the screen were absent; and (2) that immediately behind the screen, and at points of the hemispherical surface as well, the amplitude is zero, the wave being entirely cut off by the screen. This is, of course, an approximation, since at the edge of a slit, for example, the amplitude of the wave does not suddenly jump from zero to a finite value. The exact treatment is exceedingly difficult, but in the few cases for which it has been worked out, it substantiates our approximations, if the dimensions of the aperture are large compared with the wave length, showing that they lead to a correct qualitative understanding of the phenomena, though there are quantitative deviations.

To find the disturbance at P, then, we integrate over the surface, but set the integrand equal to zero, except at the openings of the

screen, obtaining

$$\varphi = \int \frac{jA}{2\lambda} \frac{1}{rr_1} e^{j\omega[t-(r+r_1)/c]}[\cos(n,r) - \cos(n,r_1)]\, da,$$

the integral being over the openings. We note that only the edges of the openings are significant, the shape of the screen away from the openings being unimportant. Now let us assume, as is almost always true in practice, that the distances r_1 and r, from source to screen and from the screen to P, are large compared with the dimensions of the holes. Then $1/rr_1$ and $[\cos(n,r) - \cos(n,r_1)]$ are so nearly constant over the aperture that we may take them outside the integral, replacing r and r_1 by mean values $\bar{r}$ and $\bar{r}_1$. If in addition we write $r + r_1$ in the exponential as $\bar{r} + \bar{r}_1 + r' + r_1'$, where r' and r_1' are the small differences between r and r_1, and their values at some mean point of the aperture, we have finally

$$\varphi = \frac{jA}{2} \frac{1}{\bar{r}\bar{r}_1}[\cos(n,\bar{r}) - \cos(n,\bar{r}_1)] e^{j\omega[t-(\bar{r}+\bar{r}_1)/c]} \int e^{-jk(r'+r_1')}\, da. \quad (4.1)$$

The whole factor outside the integral may be taken as a constant factor so that, if we are interested only in relative intensities, we may leave it out of account. We finally have a sinusoidal vibration of which the amplitudes of the components of the two phases are proportional to $C' = \int \cos k(r' + r_1')\, da$, and $S' = \int \sin k(r' + r_1')\, da$. Hence the intensity is proportional to $C'^2 + S'^2$, and our task is to compute this value.

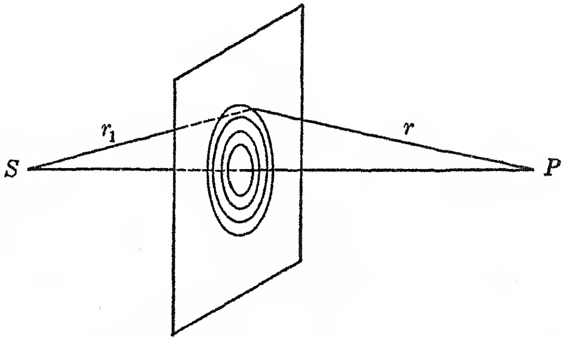


FIG. 29.—Fresnel's zones on a plane.

By using Fresnel's zones, one can see qualitatively the explanation of the diffraction fringes, particularly in Fresnel diffraction. Suppose that we join the source S and a point P with a straight line, as in Fig. 29, and consider the point of the screen cut by this line, a point

for which $r + r_1$ has a minimum value. Let us surround this point by successive closed curves in which $r + r_1$ differs from its minimum value by successive whole numbers of half wave lengths. It is not hard to see that these curves will be the intersections with the screen of a set of ellipsoids of revolution, whose foci are S and P. Hence if the line SP is approximately normal to the screen, the curves will be approximately circles. Successive zones included between successive curves will propagate light differing by a half wave length from their neighbors. Now on the screen we may imagine the pattern of zones, and also the apertures. The whole nature of the diffraction depends on what zones are uncovered, and can transmit light, and what ones are obscured by the screen. We may distinguish three cases, shown in Fig. 30:

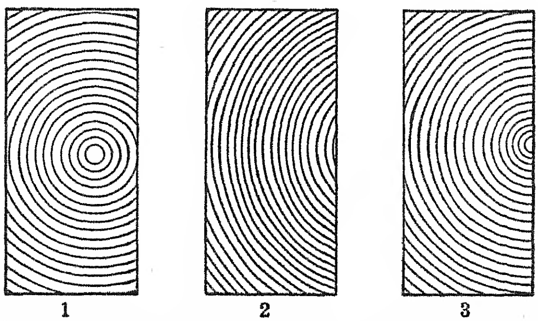


FIG. 30.—Fresnel's zones and rectangular aperture. (1) Directly in path of light. (2) In geometrical shadow. (3) On edge of shadow.

1. The center of the system of zones lies well inside the aperture. The central zone is entirely uncovered, as are a number of the others. As we get to larger zones, we shall come to one of which a small part is covered; then one that is more covered; and so on, until finally we come to one only slightly uncovered; and then the rest are entirely obscured. Now we can write our integral, as in Sec. 3, as a sum of integrals over the successive zones. As before, these contributions will decrease very gradually from one zone to the next. When we reach the zones that are obscured, the decrease will become a little more rapid, but not so much as to interfere with the argument. We can still write the whole thing as half the sum of the first and the last zones. In our case, the last zone that contributes has a negligibly small area exposed, so that it contributes practically nothing, and the whole integral is half the first zone. But this gives just the intensity we should have in the absence of the screen.

2. The center of the zone system is well behind the screen (P is in the geometrical shadow). Then the first few zones are obscured. A certain zone begins to be uncovered, until finally some zones are uncovered to a considerable extent. Large zones become obscured again, however. Thus in our sum, although there are terms different from zero, both the first and the last terms are zero, so that the sum is zero. The intensity well inside the geometrical shadow is zero.

3. The center of the zone system is near the edge of the screen. Then the first zone may be partly obscured, so that there is some intensity, but not so great as without the screen. Or the first zone may be entirely uncovered, but the next ones partly obscured. In these cases, the contributions from the successive zones may differ so much that our rule of taking the first and last terms is no longer correct. It is possible for the whole amplitude to be more than half the first zone, so that the intensity is actually greater than without the screen. As we move into the geometrical image from the shadow, it turns out that there is a periodic fluctuation, because of the uncovering of successive zones, and this explains the diffraction fringes.

Problems

1. Try to carry out exactly the integration that we did approximately by using Fresnel's zones.

2. The source is at infinity, so that a wave front is a plane. Set up Fresnel's zones, and find the breadth of the nth zone, and its area.

3. A plane wave falls on a screen in which there is a circular hole. Investigate the amplitude of the diffracted wave at a point on the axis, showing that there is alternate light and darkness either as the radius of the hole increases, or as the point moves toward or away from the screen. (*Suggestion:* The integral consists of a finite number of zones.)

4. A plane wave falls on a circular obstacle. Show that at a point behind the obstacle, precisely on the axis, there is illumination of the same intensity that we should have if the obstacle were not there. Explain why this would not hold for other shapes of the obstacle.

5. Take a few simple alternating series, such as, $1/2 - 1/3 + 1/4 - 1/5 + \cdots$, $1/2 - 1/4 + 1/6 - \cdots$, $1/2^2 - 1/3^2 + 1/4^2 - \cdots$, and so forth, and find whether our theorem about the sum of a number of terms is verified for them. In doing this, it may be necessary to start fairly well out in the series, so as to satisfy our condition that successive terms differ only slightly in magnitude.

6. Prove the statement that the boundaries of Fresnel's zones are the intersection of the screen with ellipsoids of revolution whose foci are the source and the point P. What happens to these ellipsoids as the source is removed to infinity?

7. A "zone plate" is made by constructing 20 Fresnel zones on a plane glass plate for a source at infinity and blocking off the light from every other zone. This zone plane, when illuminated at normal incidence by a plane wave, acts as a lens and produces a bright spot on its axis at a distance of 1 m from the plate.

If the incident light is monochromatic and of wave length 5,000 A, compute the radius of the zone plate and the intensity at the bright spot relative to its value in the absence of the zone plate. At what other points on the axis will there be intensity maxima?

8. Consider a line source of light and a plane observation screen parallel to and at a finite distance from the line source. The wave fronts are cylindrical. Show that the Fresnel zones on a wave front at a distance d from the source are strips parallel to the line source, and compute the angles subtended by these zones at the source. Suppose a straight-edged obstacle is inserted between the source and the screen, cutting off all the light below the plane perpendicular to the screen and passing through the light source. Discuss the variation of light intensity as a function of position on the screen, and obtain approximate expressions for the location of the maxima and minima.

CHAPTER XIV

FRESNEL AND FRAUNHOFER DIFFRACTION

In the present chapter we proceed to the mathematical discussion of Fresnel and Fraunhofer diffraction, based on the methods of Huygens' principle derived in Chap. XIII. The problems that we take up are Fresnel and Fraunhofer diffraction through a slit; Fraunhofer diffraction through a circular aperture; and the diffraction grating, an example of Fraunhofer diffraction. In Eq. (4.1) of the preceding chapter, we saw that the essential step in computing the diffraction pattern is the evaluation of the integral

$$\int e^{-jk(r+r_1)}\, da,$$

where $k = 2\pi/\lambda$, and where the integration is over the aperture of the screen, da is an element of surface in the aperture, r is the distance from the source to the element da, and r_1 the distance from the element to the point P where the field is being computed. If the incident wave is a plane wave, and the plane of the aperture is a wave front, then r is the same for all elements, and the factor e^{-jkr} can be canceled out of the integral. The remaining integral, $\int e^{-jkr_1}\, da$, represents the sum at P of the amplitudes of spherical waves of equal intensity and phase starting from all points of the aperture. It is the interference of these waves which produces the diffraction pattern.

1. Comparison of Fresnel and Fraunhofer Diffraction.—The two types of diffraction, Fresnel and Fraunhofer, arise from observing the pattern near to, or far from, the screen. Let the normal to the screen be the z axis, as in Fig. 31, and let the screen containing the aperture be at $z = 0$. The light passing through the aperture is caught on a second screen at $z = R$. Physically, the diffraction pattern has the following nature: close to the aperture, the light passes along the z axis as a column or cylinder of illumination, of cross section identical with the aperture, so that, if the screen at R is close to the aperture, the illuminated region will have the same shape as the aperture, and we speak of rectilinear propagation of the light. As R increases, however, the column of light begins to acquire fluctuations of intensity near its boundaries, so that the pattern on the screen

has fringes around the edges. This phenomenon is the Fresnel diffraction.

The size of the Fresnel fringes increases proportionally to the square root of the distance R. Thus Fig. 32 shows, in its upper diagram, the slit, parallel column of light, and parabolic lines starting from the edges of the slit, indicating the position of the outer bright fringe of the Fresnel pattern, if we are sufficiently near to the slit. As R becomes larger, the fringes become so large that there are only one or two in the pattern of the aperture, and the pattern shows but small resemblance to the shape of the aperture, though it still is of roughly

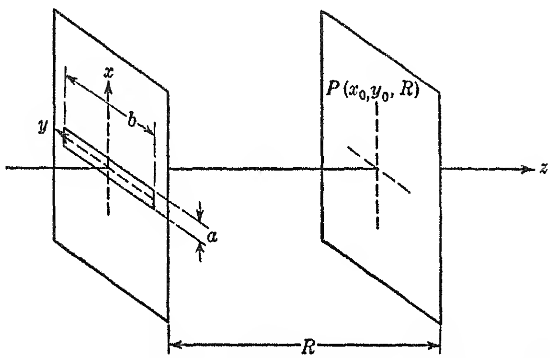


FIG. 31.—Aperture and screen for diffraction through rectangular slit.

the same dimensions. With further increase of R, we finally enter the region of Fraunhofer diffraction. Here the beam of light, instead of consisting of a luminous cylinder, resembles more a luminous cone indicated by the diverging dotted lines in the top diagram of Fig. 32. Thus the Fraunhofer pattern becomes larger and larger as R increases, being in fact proportional to R, so that we can describe it by giving the angles rather than distances between different fringes.

Often Fraunhofer diffraction is observed, not by placing the screen at a great distance, but by passing the light through a telescope focused on infinity. Such a telescope brings the light in a given direction to a focus at a given point of the field. Thus it separates the different Fraunhofer fringes, since each of these goes out from the source in a particular direction. In Fig. 32, diffraction patterns are shown indicating the transition from Fresnel to Fraunhofer diffraction. Pattern a illustrates the Fresnel pattern for one edge of an infinitely wide slit. Patterns b to g represent the actual diffraction patterns from the slit, at distances indicated in the upper diagram.

These patterns are all drawn to the same scale. They are drawn from a slit five wave lengths wide, for the sake of getting the figure on a diagram of reasonable scale. If the wave length were shorter, then for the same slit the distances would be stretched out to the

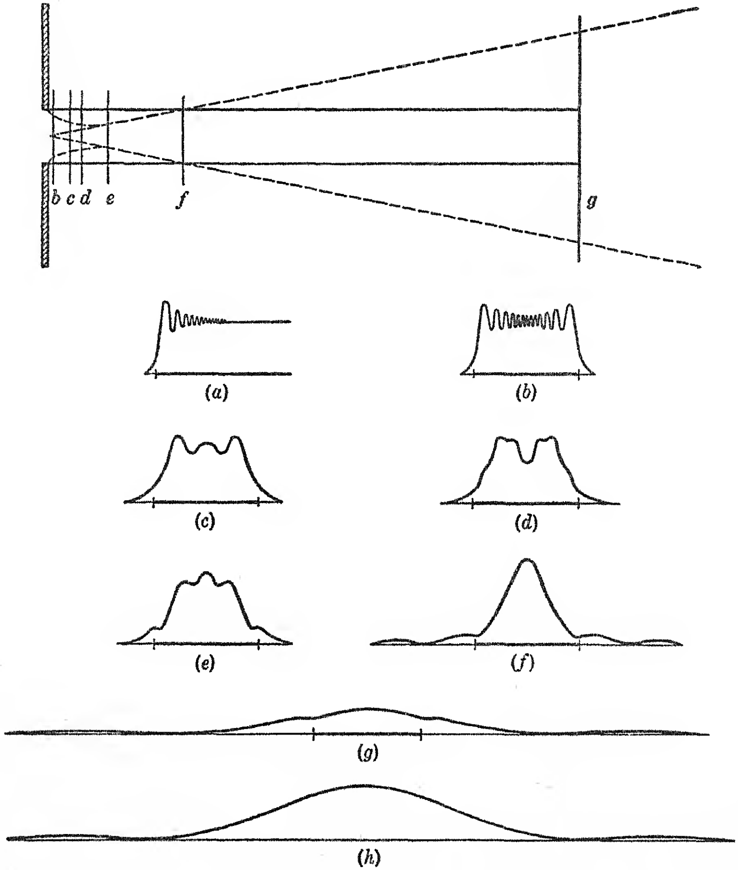


FIG. 32.—Transition from Fresnel to Fraunhofer diffraction for a slit. (a) Fresnel pattern for edge of infinitely wide slit. (b)–(g) Actual diffraction patterns from slit, at distances indicated in upper diagram. (h) Fraunhofer pattern.

right, and the Fraunhofer pattern would correspond to smaller angular deflections. This would be necessary to bring the Fresnel cases far enough from the slit so that our approximations would be really applicable. Finally, in h, we give the limiting Fraunhofer pattern, not drawn to scale.

Let coordinates in the plane of the aperture be x, y, and in the

plane of the screen at R let the coordinates be x_0, y_0, as in Fig. 31. Then, if the element of area is at x, y, 0, and the point P at x_0, y_0, R, the distance r_1 between them is

$$r_1 = \sqrt{(x_0 - x)^2 + (y_0 - y)^2 + R^2}.$$

The integration cannot be performed with this expression for r_1, and Fresnel and Fraunhofer diffraction lead to two different approximate methods of rewriting r_1, leading to different methods of evaluating the integral. We can see the relation of these two methods most clearly from Fig. 33, in which r_1 is plotted as a function of $x_0 - x$, for the special case where $y_0 - y$ is zero. The resulting curve is a hyperbola.

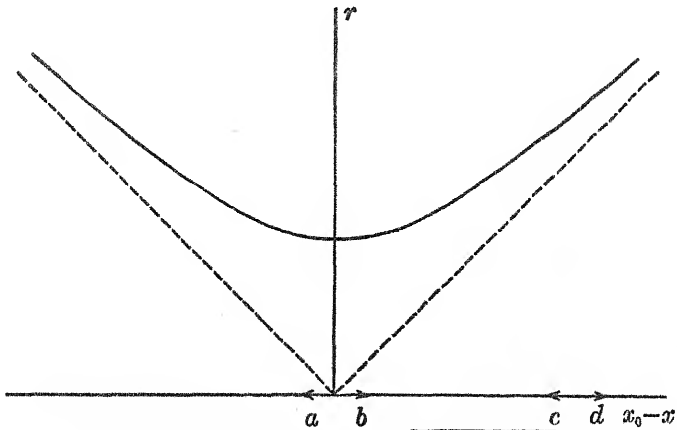


FIG. 33.—r_1 as function of $x_0 - x$. $r_1 = \sqrt{(x_0 - x)^2 + R^2}$. r_1 is the distance from a point of the aperture to a point on the screen; $x_0 - x$ is the difference between the x coordinates of the points.

Now in all ordinary cases, R is large compared with the dimensions of the aperture. That is, the range of abscissas representing the dimensions of the aperture (from $x_0 - x_1$ to $x_0 - x_2$, if x_1 and x_2 are the extreme coordinates of the aperture) is small compared with the distance R, the intercept of the hyperbola on the axis of ordinates. The two cases are now represented by the ranges ab and cd of abscissas, respectively. In the first, $x_0 - x_1$ and $x_0 - x_2$ are separately small, as well as their difference, and this means that the point P is almost straight behind the aperture, in the region where the Fresnel diffraction pattern occurs. In the second, x_0 is large, of the same order of magnitude as R, showing that we are examining the pattern at a considerable angle to the normal, as we do in the Fraunhofer case.

The two approximate methods can now be simply described from the curve: for Fresnel diffraction, we approximate the hyperbola near its minimum by a parabola; for Fraunhofer diffraction, we approximate it farther out by a straight line. In the first case, assuming R to be large compared with $(x_0 - x)$, we have by the binomial expansion

$$r_1 = R + \frac{1}{2}\frac{(x_0 - x)^2}{R} + \cdots,$$

or, including the terms in y,

$$r_1 = R + \frac{1}{2}\frac{(x_0 - x)^2 + (y_0 - y)^2}{R} + \cdots. \tag{1.1}$$

In this case, in the notation of Eq. (4.1) of Chap. XIII, we take $\bar{r} = R$, so that r' is the remaining term of (1.1). For Fraunhofer diffraction, on the other hand, we have $x_0 \gg x$. Then we can write $r_1^2 = (x_0^2 + y_0^2 + R^2) - 2(xx_0 + yy_0) + x^2 + y^2$, and we can neglect the terms $x^2 + y^2$. If we let $R_0^2 = x_0^2 + y_0^2 + R^2$, where R_0 measures the distance from the center of the aperture to the point P, we can use a binomial expansion, obtaining

$$r_1 = R_0 - \frac{xx_0 + yy_0}{R_0} - \cdots. \tag{1.2}$$

In this case we take $\bar{r} = R_0$, so that r' is the remaining term of (1.2). Letting $x_0/R_0 = l$, $y_0/R_0 = m$, the direction cosines of the direction from the center of the aperture to P, we have

$$r' = -(lx + my) \cdots,$$

involving the position on the screen only through the angles, so that we see at once that the pattern will travel outward radially from the aperture.

2. Fresnel Diffraction from a Slit.—Let the aperture be a slit, extending from $x = -(a/2)$ to $x = a/2$, and from $y = -(b/2)$ to $b/2$. We assume a to be small, b comparatively large, as in Fig. 31, so that it is a long narrow slit. Using the results of (1.1), our integral is

$$\int e^{-jkr'}\,da = \int e^{-\pi j[(x-x_0)^2+(y-y_0)^2]/R\lambda}\,da.$$

This can be immediately factored into

$$\int_{-b/2}^{b/2} e^{-\pi j(y-y_0)^2/R\lambda}\,dy \int_{-a/2}^{a/2} e^{-\pi j(x-x_0)^2/R\lambda}\,dx.$$

Since these two integrals are of the same form, we can treat just one of them. This will prove to give fringes parallel to one set of axes. The whole pattern is then simply the combination of the two sets of

fringes. The single integral, for instance, the one in x, has a real part, and an imaginary part (with sign changed), equal to

$$\int_{-a/2}^{a/2} \cos \frac{\pi(x - x_0)^2}{R\lambda}\, dx \qquad \text{and} \qquad \int_{-a/2}^{a/2} \sin \frac{\pi(x - x_0)^2}{R\lambda}\, dx. \tag{2.1}$$

It is customary in these integrals to make a change of variables: $\dfrac{(x - x_0)^2}{R\lambda} = \dfrac{u^2}{2}$. Then the integrals become $\sqrt{R\lambda/2}$ times C and S, respectively, where $C = \int_{u_1}^{u_2} \cos\frac{\pi}{2} u^2\, du$, $S = \int_{u_1}^{u_2} \sin\frac{\pi}{2} u^2\, du$, and where $u_1 = \dfrac{x_0 - a/2}{\sqrt{R\lambda/2}}$, $u_2 = \dfrac{x_0 + a/2}{\sqrt{R\lambda/2}}$. These integrals are called "Fresnel's integrals." They cannot be explicitly evaluated, but their values have been computed by series methods.

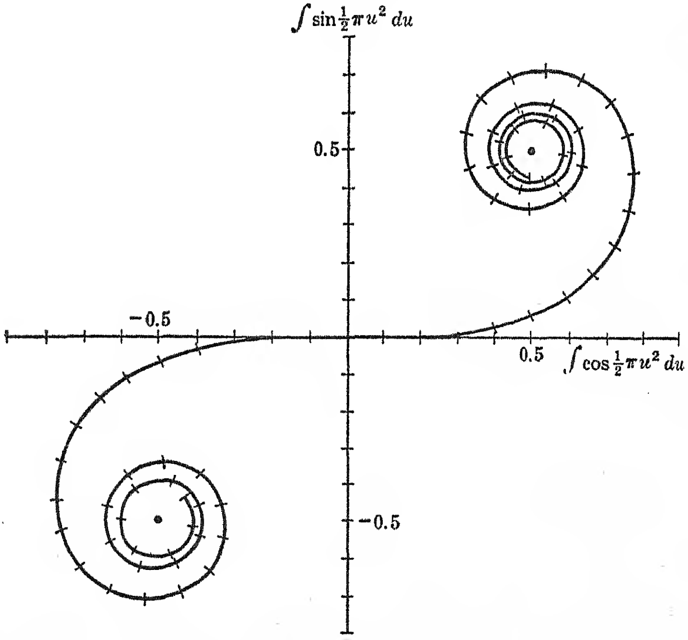


FIG. 34.—Cornu's spiral. The points of the spiral marked by cross bars correspond to increments of 0.1 unit in u.

To discuss Fresnel's integrals, let us plot the indefinite integral $\int_0^u \cos\frac{\pi}{2} u^2\, du$ as abscissa, $\int_0^u \sin\frac{\pi}{2} u^2\, du$ as ordinate, of a graph, as in Fig. 34. Then it is not hard to see that the resulting curve is a spiral,

which is known as "Cornu's spiral." To see this, we can first compute the slope. This is the differential of the ordinate, over the differential of the abscissa, or

$$\frac{\sin (\pi/2)u^2}{\cos (\pi/2)u^2} = \tan \frac{\pi}{2} u^2.$$

Thus, when u^2 increases by 4, the tangent of the curve swings around a complete cycle, and comes back to its initial value. Each point of the spiral corresponds to a particular value of u. We can show at once that the difference of u between two points is simply the length of the curve between the points. We show this for an infinitesimal element of the curve. The square of the element of length, ds^2, is equal to the sum of the squares of the differentials of abscissa and ordinate, or is $\cos^2 [(\pi/2)u^2]\, du^2 + \sin^2 [(\pi/2)u^2]\, du^2$. Hence $ds = du$, and we can integrate to get $s = u_2 - u_1$. From this fact we can make sure of the spiral nature of the curve. For one turn corresponds to an increase of u^2 by 4. That is, if u' and u'' are the values at the two ends, $u''^2 = u'^2 + 4$. This is $u''^2 - u'^2 = 4$, $(u'' - u')(u'' + u') = 4$, $u'' - u' = 4/(u'' + u')$. The difference $u'' - u'$ is, however, simply the length of the turn, so that as we go farther along, the turns become smaller and smaller, so that they eventually become zero, which is characteristic of a spiral. It is plain that the spiral is symmetric in the origin, having two points, for $u = \pm\infty$, for which it winds up on itself.

Let us take our spiral, mark on it the positions u_1 and u_2 corresponding to the limits of our integral, and draw the straight line connecting these points. The length of this line will then be proportional to the amplitude of the disturbance, and its square to the intensity. This is easy to see: the horizontal component of the line is just C, and the vertical component S, so that the square of its length is $C^2 + S^2$. Knowing this, we can easily discuss the fluctuations of intensity, as seen in Fig. 32. As x_0 changes, it is plain that u_1 and u_2 increase together, their difference remaining fixed and equal to $\dfrac{a}{\sqrt{R\lambda/2}}\cdot$ Thus essentially we have an arc of this length, sliding along the spiral, and the intensity is measured by the square of the chord between the ends of this arc.

Now when x_0 is large and negative, the arc is wound up on itself, so that its ends practically meet, and the intensity is zero. This is the situation in the shadow. As x_0 approaches the value $-a/2$, however, u_2 approaches zero, so that one end of the arc has reached the center of the figure. There are two quite different cases, depending

on whether $u_2 - u_1$ is large or small. If it is large (a large slit and relatively short distance R and small wave length), then u_1 will not be unwound much at this point. The chord will then be half the value between the two end points of the spiral, and the intensity will be one-fourth its value without the screen, and will have increased uniformly in coming out of the shadow. As we go farther along the x direction, however, the arc will begin to wind up on the other half of the spiral, producing alternations of intensity at the end of the shadow. Then for a while u_2 will be nearly at one end of the spiral, u_1 at the other, so that the intensity for some distance will be nearly constant, and the same that we should have without the slit. This is the illuminated region directly behind the slit. Finally we approach the other boundary, and u_1 commences to unwind. We then go through the same process in the opposite order.

The other quite different case comes when $u_2 - u_1$ is small, which is the case for small slit, or large wave length or distance. Then there is never a time when u_1 is on one branch of the spiral and u_2 on the other. All through the central part of the pattern, therefore, there are no fluctuations of intensity. Such fluctuations come only far to one side or the other. They come about in this way: At some places in the pattern, the arc is long enough to wind up for a whole number of turns, and the chord is practically zero, while at other places it winds up for a whole number plus a half, and the chord has a maximum. The resulting fringes are the Fraunhofer fringes which we shall now discuss by a different method.

3. Fraunhofer Diffraction from a Slit.—Using the approximation (1.2), our integral for Fraunhofer diffraction is $e^{-jkR_0}\int e^{2\pi j(lx+my)/\lambda}\,da$. The first term, as in Fresnel diffraction, contributes nothing to the relative intensities, and may be discarded. We then have

$$\int e^{2\pi j(lx+my)/\lambda}\,da,$$

as the integral whose absolute value measures the amplitude of the disturbance. Let us suppose that the aperture is the same sort of rectangle considered above, extending from $-a/2$ to $a/2$ along x, from $-b/2$ to $b/2$ along y. Then the integral is

$$\begin{aligned}\int_{-a/2}^{a/2} e^{2\pi jlx/\lambda}\,dx \int_{-b/2}^{b/2} e^{2\pi jmy/\lambda}\,dy &= \frac{(e^{\pi jla/\lambda} - e^{-\pi jla/\lambda})}{2\pi jl/\lambda}\,\frac{(e^{\pi jmb/\lambda} - e^{-\pi jmb/\lambda})}{2\pi jm/\lambda} \\ &= \frac{\sin(\pi la/\lambda)}{\pi l/\lambda}\,\frac{\sin(\pi mb/\lambda)}{\pi m/\lambda} \qquad (3.1)\end{aligned}$$

The intensity is the square of this quantity. Let us consider its dependence on the position of the point P on the screen. The coordinates of this point enter only in the expressions l, m, showing that the pattern increases in size proportionally to the distance, as if it consisted of rays traveling out in straight lines from the small aperture, rather than having an approximately constant size as with the Fresnel diffraction (see Fig. 32). When we consider the detailed behavior of the intensity as a function of the angle, we find that this can be written as $a^2 \dfrac{\sin^2 (\pi la/\lambda)}{(\pi la/\lambda)^2}$ times a similar function of m, giving a curve of the form $\dfrac{\sin^2 \alpha}{\alpha^2}$, where $\alpha = \dfrac{\pi la}{\lambda}$. This function becomes unity when $\alpha = 0$, goes to zero for $\alpha = \pi, 2\pi, 3\pi, \cdots$, with maxima of intensity approximately midway between. The maxima decrease rapidly in intensity. Thus at the points $3\pi/2$, $5\pi/2$, . . . which are approximately at the second and third maxima, the intensities are only $(2/3\pi)^2$, $(2/5\pi)^2$, . . . , or 0.045, 0.016 . . . , compared with the central maximum of 1. Let us see how the size of the fringes depends on the dimensions of the slit. The minima come for $\alpha = n\pi$, or $la/\lambda = n$, $l = n\lambda/a$. Thus we see that the greater the wave length, or the smaller the dimensions of the slit, the larger the pattern becomes.

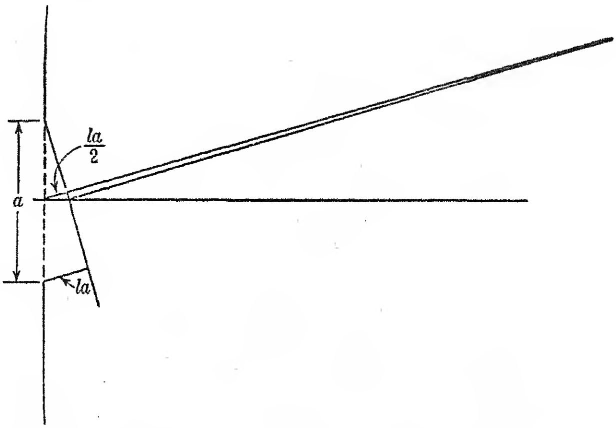


FIG. 35.—Elementary construction for Fraunhofer diffraction.

The positions of the minima can be immediately found by a very elementary argument. Assume for convenience that we are investigating the pattern at a point in the xz plane, so that $m = 0$. Then draw a plane normal to the direction l, passing through one edge of

the aperture, as in Fig. 35. This represents a wave front of the diffracted wave, just as it passes one edge of the aperture. From the geometry of the system, this wave front is a distance la from the other edge, or $la/2$ from the middle of the aperture. Now, if the distance of the middle is just a whole number of half wave lengths different from the distance from the edge, the contributions of these two points to the amplitude will just cancel, being just out of phase. The other points of one half of the aperture can all be paired against corresponding points of the other half whose contributions are just out of phase, finally resulting in zero intensity. This situation comes about when $la/2 = n\lambda/2$, where n is an integer, or $l = n\lambda/a$, the same condition found above. Since most of the intensity falls within the first minimum, and since l is the sine of the angle between the ray and the normal to the surface, we may say that by Fraunhofer diffraction the ray is spread out through an angle λ/a.

4. The Circular Aperture, and the Resolving Power of a Lens.—The problem of Fraunhofer diffraction through a circular aperture is slightly more complicated mathematically. Here we must evaluate $\int e^{2\pi j(lx+my)/\lambda}\,da$ over a circle. Let us introduce polar coordinates in the plane of the aperture, so that $x = \rho \cos \theta$, $y = \rho \sin \theta$. Further, on account of symmetry, we may take the point P to be in the xz plane, so that $m = 0$. Then, if ρ_0 is the radius of the aperture, the final result is $\int_0^{2\pi} d\theta \int_0^{\rho_0} e^{2\pi j\rho\cos\theta l/\lambda}\rho\,d\rho$. We integrate with respect to ρ by parts, obtaining for the integral

$$\int_0^{2\pi} d\theta \left[\frac{\rho_0 e^{2\pi j\rho_0\cos\theta l/\lambda}}{2\pi j\cos\theta l/\lambda} - \frac{(e^{2\pi j\rho_0\cos\theta l/\lambda} - 1)}{(2\pi j\cos\theta l/\lambda)^2}\right].$$

For the integration with respect to θ, it is necessary to expand the exponentials in series. If we do this, the integrals are in each case integrals of a power of $\cos\theta$, from 0 to 2π. These are easily evaluated, and the result, combining terms, proves to be

$$\pi\rho_0^2\left[1 - \frac{1}{2}\left(\frac{k}{1}\right)^2 + \frac{1}{3}\left(\frac{k^2}{2!}\right)^2 - \frac{1}{4}\left(\frac{k^3}{3!}\right)^2 + \frac{1}{5}\left(\frac{k^4}{4!}\right)^2 - \cdots\right],$$

where k is an abbreviation for $\pi\rho_0 l/\lambda$. If we recall the formulas for Bessel's functions, discussed in Appendix VII, we can see without difficulty that this is equal to $(\rho_0\lambda/l)J_1[2\pi\rho_0(l/\lambda)]$. It is not hard, using some of the properties of Bessel's functions, to prove this formula directly, without the use of series.

From the series, we see that the intensity has a maximum for $l = 0$, the center of the pattern. As l increases, we can see the behavior most easily from the expression in terms of Bessel's functions. Since J_1 has an infinite number of zeros, there are an infinite number of light and dark fringes. The first dark band comes at the first zero of J_1, which from tables is at $2\pi\rho_0 l/\lambda = 1.2197\pi$, $\rho_0 l/\lambda = 0.61$. The next is at $\rho_0 l/\lambda = 1.16$, and so on, with maxima between. We see that, except for a numerical factor, the pattern from a circular aperture has about the same dimensions as that from a square aperture. Thus if the side of the square were equal to the diameter of the circle, $2\rho_0$, the first dark fringe would be at $2\rho_0 l/\lambda = 1$, $\rho_0 l/\lambda = 0.5$, and the next one at 1.0.

Whenever light passes through a lens, it is not only refracted, but it has passed through a circular aperture, the size of the lens itself or of the diaphragm that stops it down, and as a result it is diffracted. Suppose, for example, that the lens is the objective of a telescope, and that parallel light falls on it, as from an infinitely small or distant star. Then after passing through the diaphragm, the light will no longer be a plane wave, but will have intensity in different directions, as shown in the last paragraph. The central maximum will have an angular radius of $0.61\lambda/\rho_0$, where ρ_0 is now the radius of the telescope objective. The resulting waves are just as if the light came from an object of this size, but passed through no diaphragm. When the telescope focuses the radiation, the result will be not a single point of light, but a circular spot surrounded by fringes, as of a star of finite diameter. For this reason, the telescope is not a perfect instrument, and one would say that its resolving power was only enough to resolve the angle $0.61\lambda/\rho_0$. This is usually taken to mean the following: If two stars had an actual angular separation of this amount, the center of the image of one star would lie on the first dark fringe of the other, and the patterns would run into each other so that they could be just resolved. We see that the larger the aperture of the telescope, or the smaller the wave length, the better is the resolution. The same general situation holds for microscope lenses.

5. Diffraction from Several Slits; the Diffraction Grating.—Suppose we have a number N of equal, parallel slits, equally spaced. Let each have a width a along the x axis, and let the spacing on centers be d, so that the centers come at $x = 0, d, \cdots (N-1)d$. Now let us find the Fraunhofer pattern. The part of the integral depending on y will be just as with the single slit, and we leave it out of account. We are left with

$$\int_{-a/2}^{a/2} e^{2\pi jlx/\lambda}\,dx + \int_{d-a/2}^{d+a/2} e^{2\pi jlx/\lambda}\,dx + \cdots + \int_{(N-1)d-a/2}^{(N-1)d+a/2} e^{2\pi jlx/\lambda}\,dx.$$

But this is, as we can immediately see, simply

$$\int_{-a/2}^{a/2} e^{2\pi jlx/\lambda}\,dx\,[1 + e^{2\pi jld/\lambda} + e^{2\pi jl2d/\lambda} + \cdots + e^{2\pi jl(N-1)d/\lambda}].$$

By the formula for the sum of a geometric series, this is

$$\int_{-a/2}^{a/2} e^{2\pi jlx/\lambda}\,dx \left(\frac{1 - e^{2\pi jlNd/\lambda}}{1 - e^{2\pi jld/\lambda}}\right).$$

Let the first term be A, the amplitude due to a single slit, which we have already evaluated. Now to find the intensity we multiply this by its conjugate, which gives

$$A^2 \frac{1 - \cos (2\pi lNd/\lambda)}{1 - \cos (2\pi ld/\lambda)} = A^2 \frac{\sin^2 (\pi lNd/\lambda)}{\sin^2 (\pi ld/\lambda)}. \tag{5.1}$$

That is, with N slits the actual intensity is that with one slit, but multiplied by a certain factor. This factor goes through zero when lNd/λ is an integer, so that l equals an integer multiplied by λ/Nd. This gives fringes with a narrow spacing, characteristic of the whole distance Nd occupied by the set of apertures, modulating the other pattern, and they are what are usually called "interference fringes," since they are due, not to diffraction from a single aperture, but to interference between different apertures. But, in addition to this, the denominator results in having these fringes of different heights. The minimum height occurs when the denominator equals unity, when the fringes are of height A^2, and the most intense fringes come when the denominator is zero. Here the ratio of numerator to denominator is evidently finite, and gives fringes of height N^2A^2. Thus the greater N is, the greater the disparity in height between the largest and smallest maximum. Evidently every Nth maximum will be high, and the high ones will be spaced according to the law $ld/\lambda = k$, an integer.

Now suppose N becomes very great, as in a diffraction grating. Then the small maxima will become so weak compared with the strong ones that only the latter need be considered. The latter will seem to consist of a set of sharp lines, with darkness between. These sharp lines come, as we have seen, at angles θ to the normal given by $k\lambda = d \sin \theta$, where k is an integer, and $\sin \theta = l$. This is the ordinary diffraction grating formula, where k is 0 for the central image, 1 for

the first-order spectrum, 2 for the second order, and so on. But we cannot entirely neglect the fact that there are other small maxima near the important ones. Thus for $ld/\lambda = k$, the intensity is N^2A^2. This comes for $lNd/\lambda = Nk$. But for $lNd/\lambda = Nk + 3/2$, we again have a secondary maximum, whose height is now

$$\frac{A^2}{\sin^2 (\pi ld/\lambda)} = \frac{A^2}{\sin^2 \pi[(Nk + 3/2)/N]} = \frac{A^2}{\sin^2 \pi[k + 3/(2N)]}.$$

Now $\sin^2\left(\pi k + \frac{3\pi}{2N}\right) = \left(\frac{3\pi}{2N}\right)^2$ approximately, if N is large, so that the height of the maximum is $4N^2A^2/9\pi^2$, or about 0.045 of the height of the highest maximum. Thus the first few secondary maxima cannot be neglected.

To get an idea of the width of the region through which the intensity is considerable, we may take the width of the first maximum. From the center to the first dark fringe, this is given by the fact that at the center $lNd/\lambda = Nk$, at the dark fringe $Nk + 1$, so that $\Delta l = \lambda/Nd$. This is closely connected with the resolving power of a grating. For a single frequency gives not a sharp set of lines, one for each order, but a set broadened by the amount we have found. Thus two neighboring frequencies, differing by $\Delta\lambda$, could not be resolved if the first minimum of one lay opposite the maximum of the other. Since $l = \lambda k/d$, this would be the case if $\Delta l = \Delta\lambda k/d = \lambda/Nd$, or if $\Delta\lambda/\lambda = 1/Nk$. The resolving power thus increases as the number of lines in the grating increases, and as the order of the spectrum increases.

Problems

1. Carry through a discussion of Fresnel diffraction from a slit, when the source is at a finite distance, directly behind the center of the slit. In what ways will the result differ from the case we have discussed?

2. Light of wave length 6,000 A falls in a parallel beam on a slit 0.1 mm broad. Work out numerical values for the intensity distribution across the slit, at three distances, first in which the Fresnel fringes are small compared with the size of the pattern, second in which they are of the same order of magnitude, and third in which they are Fraunhofer fringes. Either construct Cornu's spiral yourself, from tables of Fresnel's integrals, or use the one of Fig. 34.

3. Find the coordinates of the points at which Cornu's spiral winds up on itself. From the chord between these points, compute the intensity behind an infinitely broad slit, which essentially means no slit at all. Find whether this agrees with what you should expect it to be.

4. Prove that the maxima of the function

$$\frac{\sin^2 (\pi la/\lambda)}{(\pi la/\lambda)^2} = \frac{\sin^2 \alpha}{\alpha^2}$$

are determined by the equation $\alpha = \tan \alpha$. Find the first three solutions of this transcendental equation, and compare them with the approximate solutions $\alpha = 3\pi/2,\ 5\pi/2,\ 7\pi/2$.

5. Discuss the Fresnel diffraction pattern caused by an edge coincident with the y axis, the screen occupying one-half the xy plane. The diffraction pattern is obtained in a plane parallel to the xy plane and a distance R from it. Plot the variation of intensity of light along the x direction from a region inside the shadow to well into the directly illuminated area. Prove that the intensity of light just at the edge of the geometrical shadow is one-fourth of its value if there were no diffraction edge.

6. Evaluate the Fresnel integrals $\int_0^u \cos \frac{\pi}{2} u^2\, du$ and $\int_0^u \sin \frac{\pi}{2} u^2\, du$ in a power series. What is the range of convergence of these series?

7. Consider the Fresnel integrals in the form $\int_0^x \cos x^2\, dx$ and $\int_0^x \sin x^2\, dx$. Integrate each of these by parts according to the scheme

$$\int_0^x \cos x^2\, dx = x \cos x^2 + 2 \int_0^x x^2 \sin x^2\, dx$$

and continue this process. Show that one obtains series for the above integrals of the form

$$\int_0^x \cos x^2\, dx = S_1 \cos x^2 + S_2 \sin x^2$$

and

$$\int_0^x \sin x^2\, dx = S_1 \sin x^2 - S_2 \cos x^2$$

where S_1 and S_2 are power series in x. Find these series and their range of convergence.

8. Find a semiconvergent series for the Fresnel integrals by the following method: Write $\int_x^\infty \cos x^2\, dx = \int_x^\infty \frac{\cos x^2}{x}\, x\, dx$, and integrate by parts, repeating the process. Then write the results in the form

$$\int_x^\infty \cos x^2\, dx = \sigma_1 \cos x^2 - \sigma_2 \sin x^2 + R_c$$

where σ_1 and σ_2 are finite series of n terms each in inverse powers of x. R_c is the remainder. Show that it is given by

$$R_c = (-1)^n \frac{1 \cdot 3 \cdot 5 \cdots (4n-1)}{2^{2n}} \int_x^\infty \frac{\cos x^2\, dx}{x^{4n}}.$$

Note that the series σ_1 and σ_2 are alternating series with terms that initially decrease with increasing n but eventually increase again. Thus the best approximation will be obtained when one stops at such a value of n that the ratio of the remainder when $n + 1$ terms are retained to that for n terms is most nearly equal to unity.

Since

$$|R_c| < R_n = \frac{1 \cdot 3 \cdot 5 \cdot \cdot \cdot (4n - 1)}{2^{2n}} \int_x^\infty \frac{dx}{x^{4n}}$$

evaluate R_{n+1}/R_n, and show that the best approximation occurs for n equal about to $x^2/2$. Find the corresponding series for $\int_x^\infty \sin x^2 \, dx$.

9. Show that in a diffraction grating all the even-order spectra will be missing if the slit separation d is twice the slit width a. Which orders will be missing if $d = 3a$? Plot the intensity distribution in the diffraction pattern formed by a grating of four equally spaced slits with $d = 3a$.

APPENDIX I

VECTORS

The study of vectors and vector operations involves two branches: vector algebra, including the addition and multiplication of vectors, and the relations between the components of vectors in different coordinate systems; and vector analysis, including the differential vector operations, corresponding integral relations, and the general theorems concerning integrals. We shall treat both subdivisions in the present appendix.

Vectors and Their Components.—We shall denote a vector, a quantity having direction as well as magnitude, by bold-faced type, as **F**. Vectors are often described by giving their components along three axes at right angles, as F_x, F_y, F_z. Their mathematical relationships are conveniently stated in terms of their components. Thus their law of addition is the parallelogram law; in terms of components, this means that, if two vectors **F** and **G** have components F_x, F_y, F_z, and G_x, G_y, G_z, respectively, the components of the sum **F** + **G** are $(F_x + G_x)$, $(F_y + G_y)$, $(F_z + G_z)$, as we show graphically in Fig. 36. To multiply a vector by a constant, as C, the vector must be increased in length by the factor C, leaving its direction unchanged; this amounts to multiplying each component by the constant C, so that the components of $C\mathbf{F}$ are CF_x, CF_y, CF_z. Often a constant like C is called a "scalar," to distinguish it from a vector. A scalar is a quantity that has magnitude but not direction, a vector having both magnitude and direction.

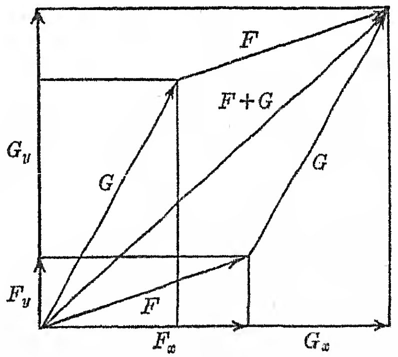


FIG. 36.—Parallelogram law for addition of vectors. The vector **F** + **G**, the diagonal of the parallelogram of sides **F**, **G**, is the vector sum of **F** and **G**. Its x component is $F_x + G_x$, its y component $F_y + G_y$.

It is often useful to write vectors in terms of three so-called "unit vectors," **i**, **j**, **k**. Here **i** is a vector of unit length, pointing along the

x axis, and similarly $\mathbf{j}$ has unit length and points along the y axis, and $\mathbf{k}$ along the z axis. Now we can build up a vector $\mathbf{F}$ out of them, by forming the quantity $\mathbf{i}F_x + \mathbf{j}F_y + \mathbf{k}F_z$. This is the sum of three vectors, one along each of the three axes; and the first, which is just the component of the whole vector along the x axis, is F_x, and the other components likewise are F_y and F_z. Thus the final vector has the components F_x, F_y, F_z, and is just the vector $\mathbf{F}$.

By the magnitude of a vector we mean its length. By the three-dimensional analogue to the Pythagorean theorem, by which the square of the diagonal of a rectangular prism is the sum of the squares of the three sides, the magnitude of the vector $\mathbf{F}$ equals $\sqrt{F_x^2 + F_y^2 + F_z^2}$. We often speak of unit vectors, vectors whose magnitude is 1. The component of a vector in a given direction is simply the projection of the vector along a line in that direction. It evidently equals the magnitude of the vector, times the cosine of the angle between the direction of the vector and the desired direction. As a special example, the component of a vector $\mathbf{F}$ along the x axis is F_x, and this must equal the magnitude of $\mathbf{F}$, times the cosine of the angle between $\mathbf{F}$ and x. If this angle is called (F,x), then we must have

$$\cos\,(F,x) = \frac{F_x}{\sqrt{F_x^2 + F_y^2 + F_z^2}}$$

with similar formulas for y and z components. The three cosines of the angles between a given direction, as the direction of the vector $\mathbf{F}$, and the three axes, are called "direction cosines," and are often denoted by letters l, m, n, so that in this case we have $l = \cos\,(F,x)$, etc. It follows immediately that $l^2 + m^2 + n^2 = 1$. We can give a simple interpretation of the direction cosines of any direction: they are the components of a unit vector in the desired direction, along the three coordinate axes.

Scalar and Vector Products of Two Vectors.—Multiplication of two vectors is a somewhat arbitrary process, governed by rules that we must postulate. It has proved to be convenient to define two entirely independent products, called the "scalar product" and the "vector product." We shall first consider the scalar product. The scalar product of two vectors $\mathbf{F}$ and $\mathbf{G}$ is denoted by $\mathbf{F}\cdot\mathbf{G}$, and by definition it is a scalar, equal to either (1) the magnitude of $\mathbf{F}$ times the magnitude of $\mathbf{G}$ times the cosine of the angle between; or (2) the magnitude of $\mathbf{F}$ times the projection of $\mathbf{G}$ on $\mathbf{F}$; or (3) the magnitude of $\mathbf{G}$ times the projection of $\mathbf{F}$ on $\mathbf{G}$. From the preceding section we see that these definitions are equivalent.

It is often useful to have the scalar product of two vectors in terms of the components along x, y, and z. We find this by writing in terms of **i**, **j**, and **k**. Thus we have

$$\begin{aligned}\mathbf{F}\cdot\mathbf{G} &= (\mathbf{i}F_x + \mathbf{j}F_y + \mathbf{k}F_z)\cdot(\mathbf{i}G_x + \mathbf{j}G_y + \mathbf{k}G_z) \\ &= (\mathbf{i}\cdot\mathbf{i})F_xG_x + (\mathbf{i}\cdot\mathbf{j})F_xG_y + (\mathbf{i}\cdot\mathbf{k})F_xG_z + (\mathbf{j}\cdot\mathbf{i})F_yG_x \\ &+ (\mathbf{j}\cdot\mathbf{j})F_yG_y + (\mathbf{j}\cdot\mathbf{k})F_yG_z + (\mathbf{k}\cdot\mathbf{i})F_zG_x + (\mathbf{k}\cdot\mathbf{j})F_zG_y + (\mathbf{k}\cdot\mathbf{k})F_zG_z.\end{aligned}$$

But, by the fundamental definition,

$$\begin{aligned}&\mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} = 1, \\ &\mathbf{i}\cdot\mathbf{j} = \mathbf{j}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{k} = \mathbf{k}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{i} = \mathbf{i}\cdot\mathbf{k} = 0.\end{aligned} \tag{1}$$

Thus

$$\mathbf{F}\cdot\mathbf{G} = F_xG_x + F_yG_y + F_zG_z. \tag{2}$$

The scalar product has many uses, principally in cases where we are interested in the projection of vectors. For example, the scalar product of a vector with a unit vector in a given direction equals the projection of the vector in the desired direction. The scalar product of a vector with itself equals the square of its magnitude, and is often denoted by F^2. The scalar product of two unit vectors gives the cosine of the angle between the directions of the two vectors. To prove that two vectors are at right angles, we need merely prove that their scalar product vanishes.

The vector product of two vectors **F** and **G** is denoted by $(\mathbf{F}\times\mathbf{G})$, and by definition it is a vector, at right angles to the plane of the two vectors, equal in magnitude to either (1) the magnitude of **F** times the magnitude of **G** times the sine of the angle between them; or (2) the magnitude of **F** times the projection of **G** on the plane normal to **F**; or (3) the magnitude of **G** times the projection of **F** on the plane normal to **G**. We must further specify the sense of the vector, whether it points up or down from the plane. This is shown in Fig. 37, where we see that **F**, **G**, and $\mathbf{F}\times\mathbf{G}$ have the same relations as the coordinates x, y, z in a right-handed system of coordinates.

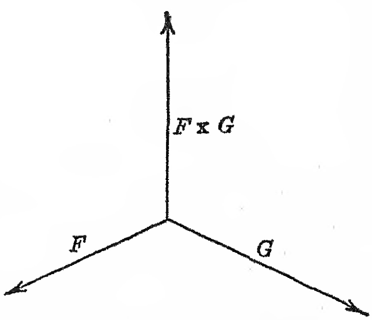


FIG. 37.—Direction of the vector product.

Another way to describe the rule in words is that, if one rotates **F** into **G**, the rotation is such that a right-handed screw turning in that direction would be driven along the direction of the vector

product. From this rule, we note one interesting fact: if we interchange the order of the factors, we reverse the vector. Thus

$$(\mathbf{F} \times \mathbf{G}) = -(\mathbf{G} \times \mathbf{F}).$$

We can compute the vector product in terms of the components, much as we did with the scalar product. Thus we have

$$\begin{aligned} \mathbf{F} \times \mathbf{G} = (\mathbf{i} \times \mathbf{i})F_xG_x + (\mathbf{i} \times \mathbf{j})F_xG_y + (\mathbf{i} \times \mathbf{k})F_xG_z \\ + (\mathbf{j} \times \mathbf{i})F_yG_x + (\mathbf{j} \times \mathbf{j})F_yG_y + (\mathbf{j} \times \mathbf{k})F_yG_z \\ + (\mathbf{k} \times \mathbf{i})F_zG_x + (\mathbf{k} \times \mathbf{j})F_zG_y + (\mathbf{k} \times \mathbf{k})F_zG_z. \end{aligned}$$

But now, as we readily see from the definition,

$$(\mathbf{i} \times \mathbf{i}) = (\mathbf{j} \times \mathbf{j}) = (\mathbf{k} \times \mathbf{k}) = 0$$

(as, in fact, the vector product of any vector with itself is zero), and

$$(\mathbf{i} \times \mathbf{j}) = -(\mathbf{j} \times \mathbf{i}) = \mathbf{k}, \qquad (\mathbf{j} \times \mathbf{k}) = -(\mathbf{k} \times \mathbf{j}) = \mathbf{i}, \qquad (\mathbf{k} \times \mathbf{i}) = -(\mathbf{i} \times \mathbf{k}) = \mathbf{j}. \quad (3)$$

Hence, rearranging terms, we have

$$\mathbf{F} \times \mathbf{G} = \mathbf{i}(F_yG_z - F_zG_y) + \mathbf{j}(F_zG_x - F_xG_z) + \mathbf{k}(F_xG_y - F_yG_x). \quad (4)$$

The Differentiation of Vectors.—We have seen that there are at least three processes of multiplication involving vectors: the multiplication of a vector by a scalar, the scalar product of two vectors, and the vector product of two vectors. In a somewhat similar way there are a number of differential operations involving vectors, all with their special uses. The simplest of these is the differentiation of a vector with respect to a scalar, such as the time, in which the x, y, and z components of the time derivative of a vector are simply the time derivative of the x, y, and z components of the vector.

The next type of differentiation that we consider is the differentiation of a scalar with respect to x, y, and z, to give a vector, the gradient, which we encountered in Chap. I, Sec. 3, in discussing the relationship between the electric field $\mathbf{E}$ and the scalar potential. The gradient of a potential φ is defined as

$$\text{grad } \varphi = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)\varphi.$$

Here we have written the gradient symbolically as the product of a vector operator, and the scalar φ. This vector operator is ordinarily

denoted by a special symbol ∇ (pronounced "del"), defined by

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}. \tag{5}$$

Whenever we use the operator ∇, we understand that the differentiations are to operate on whatever appears to the right of the operator. With this definition, we see that we may write the identity

$$\text{grad } \varphi = \nabla\varphi.$$

In some texts on vectors, the gradient is simply denoted by $\nabla\varphi$. As we see in Chap. I, Sec. 3, the physical significance of the gradient is simple. The gradient of a scalar points at right angles to the surfaces on which that function is constant, or in the direction in which the function changes most rapidly with position, and its magnitude measures the rate of change of the function in that direction. Its component in any direction, often called the "directional derivative" of the function in that direction, measures the rate of change of the function in that direction, as for instance $\partial\varphi/\partial x$ measures the rate of change of φ with x.

Using the vector operator ∇, we can now define two types of derivatives of a vector: if we have a vector $\mathbf{F}$, we can define derivatives by taking the scalar or the vector product of ∇ with the vector, resulting in $\nabla \cdot \mathbf{F}$ and $\nabla \times \mathbf{F}$. These are the quantities defined as the divergence and the curl. From the definition (5), we have

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$$

and

$$\text{curl } \mathbf{F} = \nabla \times \mathbf{F} = \mathbf{i}\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right).$$

There are also various vector operations involving second derivatives. The most familiar ones involve the operator $\nabla \cdot \nabla = \nabla^2$, which written out is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

This is the operator that we encounter often in the wave equation, Laplace's equation, etc., and that is often called the "Laplacian," for that reason. Being a scalar operator, it can be applied to either a scalar or a vector, and both forms frequently occur. Another second differential operator that is encountered often in electromag-

netic theory, though not often in mechanics, is the vector operator curl curl, applied to a vector: curl curl $\mathbf{F} = \nabla \times (\nabla \times \mathbf{F})$. We prove in a problem the useful relation

$$\text{curl curl } \mathbf{F} = \text{grad div } \mathbf{F} - \nabla^2 \mathbf{F},$$

reducing this operator to the Laplacian, which we have already met, and to grad div $\mathbf{F} = \nabla(\nabla \cdot \mathbf{F})$. This completes the list of the vector operations that are often encountered.

The Divergence Theorem and Stokes's Theorem.—There are two vector theorems involving integrals, somewhat similar in principle to the theorem regarding integration by parts in calculus, which are of great importance in vector analysis. These are the divergence theorem, sometimes known as Gauss's theorem, and Stokes's theorem. We shall now prove these theorems. The divergence theorem relates to a closed volume V in space, and the surface S that bounds it. It is assumed that there is a vector function of position $\mathbf{F}$. The theorem then states that the surface integral of the normal component of $\mathbf{F}$, over the surface S, equals the volume integral of div $\mathbf{F}$, over the volume V. That is,

$$\int_S\int F_n \, da = \int_S\int \mathbf{F} \cdot \mathbf{n} \, da = \int\int_V\int \text{div } \mathbf{F} \, dv, \tag{6}$$

where $\mathbf{n}$ is unit vector along the outer normal, so that $\mathbf{F} \cdot \mathbf{n}$ is another way of writing F_n, the component of $\mathbf{F}$ along the outer normal. To prove our theorem, we start by dividing up the volume V into thin

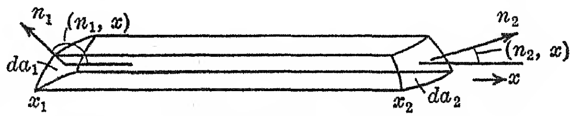


FIG. 38.—Construction for the divergence theorem.

elements bounded by planes y = constant, z = constant, as in Fig. 38. The component F_x will be a function of x along such an element, and we have obviously

$$\int_{x_1}^{x_2} \frac{\partial F_x}{\partial x} \, dx = F_x \Big|_{x_1}^{x_2}, \tag{7}$$

where x_1, x_2 are the values of x at which the element cuts through the surface S. Let $\mathbf{n}_1$, $\mathbf{n}_2$ be the outer normals at these two ends of the element, and let da_1, da_2 be the corresponding areas of surface bounding

the ends of the element. We shall have

$$\begin{aligned} -da_1 \cos (n_1,x) &= dy\, dz \\ da_2 \cos (n_2,x) &= dy\, dz \end{aligned}$$

where (n_1,x) and (n_2,x) refer to the angles between the vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ and the x axis, where the negative sign in the first equation arises because the outer normal has a negative projection on the x axis at the end x_1, and where dy, dz represent the thickness of the element along the y and z directions.

Multiplying both sides of (7) by $dy\, dz$, we then have

$$\int_{x_1}^{x_2} \frac{\partial F_x}{\partial x}\, dx\, dy\, dz = [F_{x1} \cos (n_1,x)\, da_1 + F_{x2} \cos (n_2,x)\, da_2].$$

If we carry out a summation over all elements of this type, the integral on the left will become a volume integral over V, the sum on the right will become a surface integral over S, and we have

$$\iiint_V \frac{\partial F_x}{\partial x}\, dv = \iint_S F_x \cos (n,x)\, da. \tag{8}$$

We now proceed similarly with y and z, breaking up the volume into thin elements with their axes along y and z, and obtain two other equations similar to (8). Adding them, we have

$$\begin{aligned} &\iiint_V \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \right) dv \\ &\qquad = \iint_S [F_x \cos (n,x) + F_y \cos (n,y) + F_z \cos (n,z)]\, da. \end{aligned}$$

But the integrand on the right is just the scalar product $\mathbf{F} \cdot \mathbf{n}$, so that we have proved our theorem (6). This theorem, the divergence theorem, is met in Poisson's equation, the equation of continuity, and other places, and forms the basis of Green's theorem, which is proved in Sec. 4, Chap. II.

A simplified version of our proof, though not entirely satisfactory, is instructive. This comes from proving the theorem for a volume consisting of an infinitesimal rectangular parallelepiped, bounded by x, $x + dx$, y, $y + dy$, z, $z + dz$. First we compute the surface integral of F_n over the six faces of the volume. For the face at x, F_n is $-F_x$, and the integral is $-F_x(x)\, dy\, dz$, where the value of F_x is to be computed at x. For the face at $x + dx$, F_n is F_x, and the integral is $F_x(x + dx)\, dy\, dz$. Thus the surface integral over these two faces is

approximately $(\partial F_x/\partial x)\, dx\, dy\, dz$. Adding similar contributions from the faces normal to the y and z axes, we find for the total surface integral the amount div $\mathbf{F}\, dx\, dy\, dz$, which is what the divergence theorem would give. This treatment suggests a simple physical definition for the divergence of a vector: it is the total outward flux of the vector per unit volume. To get from this form of the theorem to that involving a finite volume, we may subdivide the finite volume into infinitesimal parallelepipeds. The total flux outward of the vector through the surface of the finite volume is just equal to the sum of the fluxes outward over the surfaces of the infinitesimal volumes; for at each interior surface of separation between infinitesimal volumes, the flux outward from one volume is just balanced by the flux into its neighbor, leaving only the contributions of the exterior surfaces. The weakness of this proof lies only in the fact that a real volume cannot be built up entirely of infinitesimal parallelepipeds; around the boundary there would have to be infinitesimal volumes with surfaces inclined to the coordinate planes, which this treatment does not consider. Our earlier treatment, however, removes this objection.

Next we consider Stokes's theorem. This theorem relates to a closed line L in space, and a surface S that is bounded by L. Again we have a vector function of position $\mathbf{F}$. Stokes's theorem states that the line integral of the tangential component of $\mathbf{F}$, around L, equals the surface integral of the normal component of curl $\mathbf{F}$, over S. That is,

$$\int_L \mathbf{F} \cdot d\mathbf{s} = \int\!\!\int_S (\text{curl } \mathbf{F}) \cdot \mathbf{n}\, da, \tag{9}$$

where $d\mathbf{s}$ is a vector element of distance around the boundary, and $\mathbf{n}$ is the normal to the surface, chosen to point up if the positive direction of traversing the boundary is chosen (that is, if we go around it in a counterclockwise direction). We shall give a simplified discussion of this theorem, similar to the second method we used for handling the divergence theorem. Let us subdivide the surface S into a set of small approximately rectangular areas. We shall prove the theorem for each of the separate areas, and then combine them. As we see from Fig. 39, the contributions to the line integrals from the interior boundaries will cancel, so that the sum of all the line integrals will equal the line integral around the perimeter. The sum of the

FIG. 39.—Surface for discussing Stokes's theorem.

surface integrals over the separate rectangles will clearly equal the surface integral over the whole surface. Thus by adding the separate contributions we arrive at the theorem (9), which we wish to prove. The only reservation is that the line integral is computed for a saw-tooth type of curve approximating the actual boundary. It is not hard to show, though we shall omit it, that the line integral $\int \mathbf{F} \cdot d\mathbf{s}$ over a saw-tooth curve approximating the actual curve sufficiently smoothly differs by a negligible amount from the integral over the actual curve, so that we can construct a rigorous proof of the theorem without trouble.

We must then prove our theorem (9) for a small rectangular area. Let us choose the x and y axes to point along the sides of the rectangle, the z axis along the normal $\mathbf{n}$; if we can prove the result in this coordinate system, it will have to hold in any other coordinate system as well. The rectangle is considered to be bounded by the values x, $x + dx$, y, $y + dy$, as in Fig. 40. The surface integral on the right of (9) is then $(\partial F_y/\partial x - \partial F_x/\partial y)\, dx\, dy$. Let us next compute $\int \mathbf{F} \cdot d\mathbf{s}$ for the element of area. It is evidently

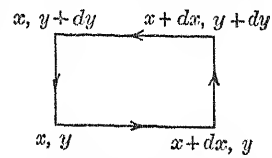


FIG. 40.—Circuit for proving Stokes's theorem.

$$F_x(x,y)\, dx + F_y(x + dx,y)\, dy - F_x(x,y + dy)\, dx - F_y(x,y)\, dy = \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy,$$

if we go around so as always to keep the surface on the left. Thus the theorem (9) is true for the infinitesimal rectangular surface, and by the argument above it then holds for a finite surface as well.

The physical meaning of the curl is easily seen from Stokes's theorem: it is the line integral of the tangential component of the vector around an area, per unit area. A vector field has a curl when its lines of force close on themselves, like the magnetic lines around a wire carrying an electric current, or like the lines of flow in a whirlpool-like motion of a fluid. A simple example, which makes the significance of the curl very clear, is the vector velocity associated with the rigid rotation of a body. Let a body have an angular velocity ω about the z axis. Then the linear velocity at point x, y, z is given by $v_x = -\omega y$, $v_y = \omega x$, $v_z = 0$. We compute the curl, and find at once that it is along the z direction, has a constant value independent of position, and is of magnitude 2ω. Thus it is simply twice the angular velocity vector of the rotating motion.

Problems

1. Find the angle between the diagonal of a cube and one of the edges. (*Hint:* Regard the diagonal as a vector **i** + **j** + **k**.)

2. Given a vector **i** + 2**j** + 3**k**, and a second **i** − 2**j** + a**k**, find a so that the two vectors are at right angles to each other.

3. Prove that $lx + my + nz = k$, where l, m, n, k are constants, and $l^2 + m^2 + n^2 = 1$ is the equation of a plane whose normal has the direction cosines l, m, n, and whose shortest distance from the origin is k.

4. Prove that $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B})$, where **A**, **B**, **C** are any vectors. Show that these are equal to the determinant

$$\begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}$$

5. Prove that $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B})$, where **A**, **B**, **C** are any vectors.

6. Prove that div $a\mathbf{F} = a$ div $\mathbf{F} + (\mathbf{F} \cdot \text{grad } a)$, where a is a scalar, **F** a vector.

7. Prove that curl $a\mathbf{F} = a$ curl $\mathbf{F} + [(\text{grad } a) \times \mathbf{F}]$, where a is a scalar, **F** a vector.

8. Prove that div $(\mathbf{F} \times \mathbf{G}) = (\mathbf{G} \cdot \text{curl } \mathbf{F}) - (\mathbf{F} \cdot \text{curl } \mathbf{G})$, where **F**, **G** are vectors.

9. Prove that div curl **F** = 0, where **F** is any vector.

10. Prove that curl curl **F** = grad div **F** − ∇^2**F**, where **F** is any vector.

11. Prove that curl grad a = 0, where a is any scalar.

APPENDIX II

UNITS

One of the complications of electromagnetic theory is the fact that a number of different systems of units are in common use. There are two so-called "absolute" cgs systems, both based on the centimeter as unit of distance, gram as unit of mass, and second as unit of time. These are the electrostatic and electromagnetic units, respectively (often abbreviated esu and emu). There is a combination of these, in which certain quantities are denoted in the electrostatic and certain others in the electromagnetic units, which is called the "Gaussian," or "mixed," units. There are the so-called "practical units," based largely on the electromagnetic units. There is the system used in this volume, the so-called mks system, in which, by using practical units for the electromagnetic quantities, but by using meters as units of distance, and kilograms as units of mass, we convert the practical system into an absolute system (that is, into a system in which various numerical constants can be set equal to unity, so that they drop out of the equations). Finally, in all these systems, we have a choice of so-called "rationalized" and "unrationalized" units, differing by factors of 4π, in which the unrationalized units give simpler formulas when we are dealing with problems of spherical symmetry, as the force between two point charges (where the force field is spherically symmetrical), while the rationalized units give simpler formulas when we are dealing with problems with rectangular symmetry, as the capacity of a parallel-plate condenser, or the properties of a plane wave. Of all these systems, the unrationalized Gaussian units, and the rationalized mks units, have the widest use and are the most valuable. In this appendix we take up the relations among these units, and the definitions of all the types of units.

We start with the unrationalized electrostatic units. Here the fundamental definition is that of the unit charge. It is so chosen as to eliminate the constant in Coulomb's law, which we have stated in mks units in Eq. (2.2), Chap. I. That is, in esu, the force in dynes between two unit charges is given by

$$F = \frac{qq'}{r^2},$$

where r is in centimeters; or unit charge is that charge which, placed at a distance of 1 cm from an equal charge, acts on it with a force of 1 dyne. The corresponding rationalized electrostatic units would be defined by the statement $F = qq'/4\pi r^2$, so that unit charge would be that which, placed at a distance of 1 cm from an equal charge, acts on it with a force of $1/4\pi$ dynes. That is, the unit charge in rationalized units is $1/\sqrt{4\pi}$ times as great as in unrationalized units. We shall not, however, consider rationalized esu further.

In unrationalized esu, we define unit current as that in which unit charge passes any point of the circuit per second. We define unit electric field $\mathbf{E}$ from the force equation, which we have given in Eq. (1.1), Chap. I, for mks units. That is, we assume that

$$\mathbf{F} = q\mathbf{E}, \tag{2}$$

for a static field, so that unit field is that in which unit charge is acted on by a force of 1 dyne. Electrostatic unit of potential is that potential difference which will impart to unit charge an energy of 1 erg. It is clear from this that the field of a charge q, at a distance r, has a magnitude q/r^2, and a scalar potential q/r, in unrationalized esu. The units of $\mathbf{D}$, in unrationalized esu, are determined by the arbitrary assumption that $\mathbf{D} = \mathbf{E}$ in empty space, in these units, so that the units of $\mathbf{D}$ are the same as those of $\mathbf{E}$. If now we have a field

$$E = D = \frac{q}{r^2}$$

from a charge q, the outward flux of $\mathbf{D}$ can be found at once by considering a sphere, and is $4\pi q$. Thus Gauss's law becomes

$$\int \mathbf{D} \cdot \mathbf{n}\, da = 4\pi \sum_i q_i = 4\pi \int \rho\, dv, \tag{3}$$

and the corresponding differential expression is

$$\operatorname{div} \mathbf{D} = 4\pi\rho. \tag{4}$$

With the formulas relating charge and field, we can find the capacity of a condenser in esu, where of course capacity is defined as the charge per unit potential difference. For a parallel-plate condenser of area A sq cm, with a separation of d cm between the plates, and empty space between, the capacity proves to be

$$C = \frac{A}{4\pi d}. \tag{5}$$

We note that the capacity, being an area divided by a length, has the dimension of length in these units; it is sometimes expressed in centimeters. We also note that the factor 4π, which did not appear in Coulomb's law (1) in unrationalized units, has appeared in the formula for the capacity of a parallel-plate condenser, just the reverse of the situation with rationalized units. If now the condenser is filled with a dielectric, the capacity will be κ_e times as great, where κ_e, the dielectric constant, being a dimensionless quantity, is the same in all systems of units. The situation is complicated by the fact, however, that in unrationalized esu the quantity κ_e is often denoted by ϵ, though with a quite different meaning for ϵ from what we have in our mks units. The capacity of a parallel-plate condenser filled with a dielectric, in unrationalized esu, is then

$$C = \frac{\kappa_e A}{4\pi d} \text{ or } \frac{\epsilon A}{4\pi d}. \tag{6}$$

It will be noted that this is related to the value $\epsilon_0 A/d$, given in Sec. 2, Chap. II, for the mks units, by having ϵ in place of ϵ_0 (since the dielectric is not equivalent to free space), and by being divided by 4π. Similarly any of the other expressions for capacities of condensers of various shape, given in that section, can be converted to unrationalized esu by changing ϵ_0 to ϵ, and by dividing by 4π.

A dielectric of course derives its properties from the dipoles within it. The potential of a dipole of moment m (which, of course, is the product of the charge, in esu, times the separation of the opposite charges, in centimeters) is $m \cos \theta/r^2$, instead of the value

$$\frac{m \cos \theta}{4\pi\epsilon_0 r^2}$$

which we have in mks units. We set up a polarization vector **P**, equal to the dipole moment per unit volume in a dielectric, just as we did in mks units. As before, we find the volume charge density arising from the polarization to be given by the relation $\operatorname{div} \mathbf{P} = -\rho'$. Assuming, as in Chap. IV, that **E** is the field arising from all charge, real and polarized, we have $\operatorname{div} \mathbf{E} = 4\pi(\rho + \rho') = 4\pi\rho - 4\pi \operatorname{div} \mathbf{P}$, $\operatorname{div} (\mathbf{E} + 4\pi\mathbf{P}) = 4\pi\rho$. Thus, comparing with (4), we see that

$$\kappa_e = 1 + \frac{4\pi P}{E} = 1 + 4\pi\chi_e, \tag{7}$$

where χ_e, the susceptibility, in the unrationalized units, is clearly $1/4\pi$ times as great as in rationalized units, as in Eq. (2.3), Chap. IV.

Taking account of the different definition of the dipole moment, we see that, in place of Eq. (1.3), Chap. IX, giving the dielectric constant of a material containing polarizable molecules, we have

$$\kappa_e = 1 + 4\pi \sum_k \frac{N_k e^2/m}{\omega_k^2 - \omega^2 + j\omega g_k}, \tag{8}$$

differing from the formula in mks units by a factor $4\pi\epsilon_0$ in the second term.

We have now discussed most of the electrostatic relations in unrationalized esu. One additional quantity is the conductivity, which is defined by Ohm's law $\mathbf{J} = \sigma\mathbf{E}$, in terms of the esu definitions of current density and electric field. When we come to magnetic quantities, it is less convenient to use esu. We can do so, by defining a magnetic field in terms of the force on a current element, but such magnetic units are seldom used, and we shall not discuss them here.

We next consider the unrationalized electromagnetic units. These are based on the law of force between two current elements as a starting point. Just as the esu of charge is chosen so as to eliminate the constant in Coulomb's law, so the emu of current is chosen so as to eliminate the constant in Eq. (1.4) of Chap. V, giving the force between two current elements. Thus in emu we have for that equation

$$d\mathbf{F} = i_1 i_2 \frac{[d\mathbf{s}_1 \times (d\mathbf{s}_2 \times \mathbf{r})]}{|\mathbf{r}|^3}. \tag{9}$$

That is, unit current in the electromagnetic system is that current which, flowing in 1 cm length of wire, acts on a similar current 1 cm away, appropriately oriented, with a force of 1 dyne. Having defined unit current, we can define unit charge in the electromagnetic system as the charge crossing a given point per second, when unit current is flowing. We can define the emu of potential as the potential difference that will impart to 1 emu of charge an energy of 1 erg. We thus note that we can get definitions of electrostatic quantities in the electromagnetic system of units. It is then a question of experiment (involving in principle the measurement of forces between charges, and forces between currents, and comparing the charges and currents) to find the relation between the electrostatic and electromagnetic units of charge, current, and potential difference. It is found that

$$\begin{aligned} 1 \text{ emu of charge} &= 2.998 \times 10^{10} \text{ esu of charge} \\ 1 \text{ emu of current} &= 2.998 \times 10^{10} \text{ esu of current} \\ 2.998 \times 10^{10} \text{ emu of potential} &= 1 \text{ esu of potential.} \end{aligned} \tag{10}$$

The quantity 2.998×10^{10} appearing in the ratio of units is that which proves to express the velocity of light in centimeters per second, and for most purposes it can be approximated by 3×10^{10} cm/sec.

The magnetic induction **B** is determined in the electromagnetic system in such a way that the force equation $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ or force per unit volume $= (\mathbf{J} \times \mathbf{B})$ will hold, without additional constants. Thus unit magnetic induction, in emu, is that induction in which a unit current flowing in a wire 1 cm long at right angles to the magnetic field is acted on by a force of 1 dyne. This emu of magnetic induction is almost the only one of the absolute units that is in common use: it is the gauss. We then find that the induction arising from an element of current is

$$d\mathbf{B} = \frac{i\,\mathbf{ds} \times \mathbf{r}}{|\mathbf{r}|^3}, \tag{11}$$

the appropriate form of the Biot-Savart law for emu. Using this law, we can find the value of **B** arising from different forms of stationary currents. Since (11) is related to Eq. (1.3) of Chap. V, the corresponding formula in mks units, by lacking the factor $\mu_0/4\pi$, it is clear that all the formulas for **B** in Secs. 2 and 3 of Chap. V can be converted to emu by multiplication by $4\pi/\mu_0$. Thus the value of **B** at a distance R from an infinite straight wire carrying a current of i (in emu) is $2i/R$, and the value of **B** inside a solenoid with n turns per unit length is $4\pi ni$. The solid angle Ω intercepted by a wire carrying a current of i emu can still be used as a scalar potential, and **B** is given by

$$\mathbf{B} = -i \operatorname{grad} \Omega. \tag{12}$$

If we define the strength of a magnetic dipole as the current, in emu, times the area, the scalar potential, whose negative gradient gives **B**, is $m \cos \theta/r^2$. The vector potential **A** still is related to **B** by the equation $\mathbf{B} = \operatorname{curl} \mathbf{A}$, where now **A** is determined by the equation

$$\mathbf{A} = \int \frac{\mathbf{J}}{r}\, dv, \tag{13}$$

instead of by Eq. (6.4), Chap. V.

From (12), we can see at once that Ampère's law for empty space is

$$\int \mathbf{B} \cdot \mathbf{ds} = 4\pi \Sigma i, \tag{14}$$

or, in differential form,

$$\operatorname{curl} \mathbf{B} = 4\pi \mathbf{J}. \tag{15}$$

In a magnetic medium, however, we have a magnetization current $\mathbf{J}'$

arising from magnetic polarization. Defining magnetic moment as we have just done, we can set up a magnetization vector **M** as the total magnetic polarization per unit volume, and we find, as in Eq. (1.2), Chap. VI, that

$$\mathbf{J}' = \text{curl } \mathbf{M}. \tag{16}$$

We then find, as in Chap. VI, that in Ampère's law in a magnetic medium, we must include the effects of all currents, both real currents and those arising from polarization, so that (15) is replaced by

$$\text{curl } \mathbf{B} = 4\pi(\mathbf{J} + \mathbf{J}') = 4\pi\mathbf{J} + 4\pi \text{ curl } \mathbf{M},$$
$$\text{curl } (\mathbf{B} - 4\pi\mathbf{M}) = 4\pi\mathbf{J}.$$

We then define a magnetic field **H**, by the equation

$$\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M},$$

in terms of which Ampère's law becomes

$$\text{curl } \mathbf{H} = 4\pi\mathbf{J}. \tag{17}$$

We see that this definition is so set up that $\mathbf{B} = \mathbf{H}$ in a nonmagnetic medium. The relation between **B** and **H** may be expressed in the forms

$$\mathbf{B} = \mathbf{H} + 4\pi\mathbf{M} = \kappa_m\mathbf{H}, \qquad \kappa_m = 1 + 4\pi\chi_m, \quad \text{where } \mathbf{M} = \chi_m\mathbf{H}. \tag{18}$$

These relations are clearly analogous to the corresponding electrostatic relations (7) and (8).

Faraday's law, in unrationalized emu, takes the forms

$$\int \mathbf{E} \cdot \mathbf{ds} = -\frac{d}{dt}\int \mathbf{B} \cdot \mathbf{n}\, da, \qquad \text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{19}$$

To see that these formulas, without additional constants, give the correct value, where **E** and **B** are expressed in emu, we may use the familiar derivation of Faraday's law from the energy principle, a derivation that we did not give in the body of the text. In this derivation, the flux through a circuit is changed by deforming the circuit, keeping the value of **B** constant in time, rather than by changing **B** with a constant circuit. The force exerted by the magnetic field on the element of length **ds** of wire carrying current i is $i(\mathbf{ds} \times \mathbf{B})$, and the work done by this field, if the wire is displaced by an amount **du**, is $i\int(\mathbf{ds} \times \mathbf{B}) \cdot \mathbf{du} = i\int\mathbf{B} \cdot (\mathbf{du} \times \mathbf{ds})$. The quantity $\mathbf{du} \times \mathbf{ds}$ is a

vector equal in magnitude to the increment of area of the circuit, and pointing along the normal $\mathbf{n}$ to the surface spanning the circuit. Thus the work done, as computed above, may be rewritten $id(\int \mathbf{B} \cdot \mathbf{n}\, da)$. To do this external work, the current flowing in the wire must have lost a compensating amount of energy, which means that an emf $\int \mathbf{E} \cdot d\mathbf{s}$ must have existed, satisfying the relation

$$id\left(\int \mathbf{B} \cdot \mathbf{n}\, da\right) = -i\, dt \int \mathbf{E} \cdot d\mathbf{s},$$

from which Faraday's law follows at once. From Faraday's law, and the Biot-Savart law (11), we see at once that Eq. (2.2) of Chap. VII, for the coefficient of self- or mutual induction, is written in unrationalized emu in the form

$$L \text{ or } M = \int\int \frac{d\mathbf{s} \cdot d\mathbf{s}'}{r}, \tag{20}$$

from which we see that the formulas of Chap. VII for inductances can be converted to emu by dividing by the factor $\mu_0/4\pi$.

We have now seen how to express the electrical quantities in unrationalized esu, and the magnetic and electrical quantities in unrationalized emu. When we combine these, to get Maxwell's equations, it is customary not to use either of these systems of units consistently, but to combine them into the Gaussian units. In these units, we use emu for $\mathbf{B}$ and $\mathbf{H}$, but esu for all other quantities, including current and charge. The electrostatic equations will then be as in esu, but the equations involving magnetic quantities will be changed. We have the relation (10) between esu and emu of charge, current, potential, and correspondingly of electric field. If we let

$$c = 2.998 \times 10^{10} \text{ cm/sec}, \tag{21}$$

we see that $\mathbf{J}$, expressed in emu, is equal to $1/c$ times $\mathbf{J}$ expressed in esu, while $\mathbf{E}$ expressed in emu is equal to c times $\mathbf{E}$ expressed in esu. Thus Maxwell's equations in Gaussian units become

$$\begin{aligned} \text{curl } \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}, & \text{div } \mathbf{B} &= 0 \\ \text{curl } \mathbf{H} &= \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} + 4\pi\frac{\mathbf{J}}{c}, & \text{div } \mathbf{D} &= 4\pi\rho, \end{aligned} \tag{22}$$

where we have included the displacement-current term, whose correctness may be inferred from the fact that it satisfies the equation of continuity, $\text{div } \mathbf{J} + \partial\rho/\partial t = 0$. These equations are to be supple-

mented by the constitutive equations,

$$\mathbf{D} = \kappa_e \mathbf{E}, \qquad \mathbf{B} = \kappa_m \mathbf{H}, \tag{23}$$

where κ_e and κ_m are as given in (7), (8), and (18), and where in Gaussian units they are often denoted by ϵ and μ, respectively. Ohm's law is still stated in the form $\mathbf{J} = \sigma \mathbf{E}$, where the value of the conductivity σ is the same in Gaussian units that it is in esu.

From Maxwell's equations in Gaussian units, we can proceed to introduce the potentials, set up wave equations, and consider energy density and the Poynting vector. For the potentials we have

$$\mathbf{B} = \text{curl } \mathbf{A}, \qquad \mathbf{E} = -\text{ grad } \varphi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}, \tag{24}$$

in which φ and $\mathbf{A}$ are subject to the relation

$$\text{div } \mathbf{A} + \frac{\kappa_e \kappa_m}{c}\frac{\partial \varphi}{\partial t} = 0. \tag{25}$$

Then we find that the potentials satisfy the following equations, if κ_e and κ_m are constants independent of position:

$$\begin{aligned} \nabla^2 \varphi - \frac{\kappa_e \kappa_m}{c^2}\frac{\partial^2 \varphi}{\partial t^2} &= -\frac{4\pi\rho}{\kappa_e} \\ \nabla^2 \mathbf{A} - \frac{\kappa_e \kappa_m}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} &= -4\pi\kappa_m \frac{\mathbf{J}}{c}. \end{aligned} \tag{26}$$

Similar wave equations are satisfied by the components of electric and magnetic field. They show that, in a region containing no charge or current, waves will be propagated with a velocity $c/\sqrt{\kappa_e \kappa_m}$, verifying our earlier statement that c represents the velocity of light in empty space. In a conducting medium in which $\mathbf{J} = \sigma \mathbf{E}$, and in which the charge density is zero, we have instead of (25)

$$\text{div } \mathbf{A} + \frac{4\pi\kappa_m \sigma \varphi}{c} + \frac{\kappa_e \kappa_m}{c}\frac{\partial \varphi}{\partial t} = 0, \tag{27}$$

and instead of (26)

$$\nabla^2 \varphi - \frac{4\pi\kappa_m \sigma}{c^2}\frac{\partial \varphi}{\partial t} - \frac{\kappa_e \kappa_m}{c^2}\frac{\partial^2 \varphi}{\partial t^2} = 0, \tag{28}$$

with a similar equation for $\mathbf{A}$.

When we consider plane wave solutions of the wave equation, we find that in empty space $\mathbf{E}$ (in esu) is numerically equal to $\mathbf{H}$ (in emu) In a medium in which κ_e and κ_m are different from unity,

$\sqrt{\kappa_e}\,\mathbf{E}$ is numerically equal to $\sqrt{\kappa_m}\,\mathbf{H}$. In a conducting medium, propagating a wave with angular frequency ω, κ_e is replaced by

$$\kappa_e - \frac{4\pi j\sigma}{\omega},$$

so that $\sqrt{\kappa_e - 4\pi j\sigma/\omega}\, E$ is equal to $\sqrt{\kappa_m}\, H$, and the complex index of refraction, instead of being $\sqrt{\kappa_e \kappa_m}$, equals $\sqrt{\kappa_m(\kappa_e - 4\pi j\sigma/\omega)}$. From these formulas, the properties of reflection from conducting surfaces, skin depth, etc., can easily be found, but it should be remembered that in Gaussian units we must use σ in esu.

From Maxwell's equations we can prove Poynting's theorem, which proves to be

$$\operatorname{div}\left[\frac{c}{4\pi}(\mathbf{E}\times\mathbf{H})\right] + \frac{\partial}{\partial t}\left[\frac{1}{8\pi}(\kappa_e E^2 + \kappa_m H^2)\right] = -\mathbf{E}\cdot\mathbf{J}. \tag{29}$$

From this we infer that Poynting's vector, in unrationalized Gaussian units, is $(c/4\pi)(\mathbf{E}\times\mathbf{H})$, and the electric- and magnetic-energy densities are $\kappa_e E^2/8\pi$ and $\kappa_m H^2/8\pi$, respectively. We note from our relation between the magnitudes of E and H, mentioned above, that the magnetic- and electric-energy densities in a plane wave are equal to each other. Applying this result to a spherical solution of the wave equation, we find that the rate of radiation from a dipole of amplitude M, angular frequency ω, is

$$\int S\, da = \frac{\omega^4 M^2}{3c^3}, \tag{30}$$

analogous to Eq. (4.2), Chap. XII. Similarly the scattering cross section of a scattering electron, analogous to (5.3), Chap. XII, is

$$\sigma = \frac{8\pi}{3}\frac{\omega^4}{(\omega_0^2 - \omega^2)^2 + (\omega g)^2}\left(\frac{e^2}{mc^2}\right)^2. \tag{31}$$

We have now indicated the equivalent, in unrationalized Gaussian units, of many of the most important formulas that have been encountered in mks units in the body of the text. Next we consider the practical units. These are based on the emu, but differ from them by arbitrary powers of 10, chosen to make the practical units of convenient size. The two most fundamental units are the ampere and the volt. These are arbitrarily defined by the following equations:

$$\begin{aligned} 1\ \text{ampere} &= 10^{-1}\ \text{emu of current} \\ 1\ \text{volt} &= 10^{8}\ \text{emu of potential.} \end{aligned} \tag{32}$$

It follows from this that in the practical system the rate of working, or power, when 1 amp flows through a potential difference of 1 volt, is 10^7 ergs/sec, or 1 joule/sec, or 1 watt, so that power in the practical system is measured in watts. The coulomb is of course defined in terms of the ampere, and the unit of electric field is ordinarily the volt per centimeter. Using (10), we see that

$$\begin{aligned} 1 \text{ esu of charge} &= \frac{1}{2.998 \times 10^9} \text{ coulombs}, \\ 1 \text{ esu of potential} &= 299.8 \text{ volts}. \end{aligned} \tag{33}$$

We see by dividing that

$$\begin{aligned} 1 \text{ ohm} &= 10^9 \text{ emu of resistance} \\ &= \frac{1}{(2.998)^2 \times 10^{11}} \text{ esu of resistance}. \end{aligned} \tag{34}$$

From Coulomb's law (1) we now see that the force between two charges q and q', expressed in coulombs, at a distance of separation r, measured in centimeters, is $F = (2.998)^2 \times 10^{18} qq'/r^2$ dynes or $(2.998)^2 \times 10^{13} qq'/r^2$ newtons. If r is measured in meters, the force is

$$F = (2.998)^2 \times 10^9 \frac{qq'}{r^2} \text{ newtons}. \tag{35}$$

This is identical with Eq. (2.2) of Chap. I, $F = qq'/4\pi\epsilon_0 r^2$, so that $\epsilon_0 = 1/[4\pi \times (2.998)^2 \times 10^9] = 8.85 \times 10^{-12}$, as in Eq. (2.1) of Chap. I. We note that (2.2) of Chap. I, which is the basis of our treatment of electrostatics in rationalized mks units, nevertheless uses as a unit the coulomb, which is an unrationalized unit, being based on the unrationalized emu. The factor 4π which appears in the definition of ϵ_0 compensates for this.

Similarly from the law (9) of force between two current elements, we see that the force between lengths ds_1 and ds_2 of wire carrying currents of i_1 and i_2, in amperes, at distance r, is

$$\begin{aligned} \mathbf{dF} &= 10^{-2} i_1 i_2 \frac{[\mathbf{ds}_1 \times (\mathbf{ds}_2 \times \mathbf{r})]}{|\mathbf{r}|^3} \quad \text{dynes} \\ &= 10^{-7} i_1 i_2 \frac{[\mathbf{ds}_1 \times (\mathbf{ds}_2 \times \mathbf{r})]}{|\mathbf{r}|^3} \quad \text{newtons}. \end{aligned} \tag{36}$$

These formulas hold whether distances are in centimeters or in meters, since the dimensions of length cancel. Thus we verify Eq. (1.4) of Chap. V, in which we see that μ_0 must be defined as in Eq. (1.2), Chap. V. When we come to magnetic fields, we find that the unit in

actual use is the gauss, which as we have seen is the emu of magnetic induction **B**. If then we have a current i, in amperes, flowing in a length **ds** of wire, in centimeters, in a magnetic induction **B**, in gausses, the force **dF**, in dynes, is

$$\mathbf{dF} = 10^{-1}i(\mathbf{ds} \times \mathbf{B}) \qquad \text{dynes.}$$

Similarly if **ds** is in meters, **dF** in newtons, we have

$$\mathbf{dF} = 10^{-4}i(\mathbf{ds} \times \mathbf{B}) \qquad \text{newtons.}$$

It is evident that to eliminate the numerical factor 10^{-4} in the mks system we must use a unit of **B** 10^4 times as large as the gauss. This unit is the weber per square meter.

The weber is a unit of flux, not of magnetic induction. It is a practical unit, based on Faraday's law: a time rate of change of flux of 1 weber/sec through a circuit induces an emf of 1 volt in that circuit. We recall, however, that a time rate of change of flux of 1 gauss/sec through an area of 1 sq cm induces an emf of 1 emu. Since 1 volt = 10^8 emu, this means that it requires a change of flux of 10^8 gausses per second flowing through an area of 1 sq cm, or of 10^4 gausses/sec flowing through 1 sq m, to induce 1 volt. In other words,

$$\begin{aligned} 1 \text{ weber/sq cm} &= 10^8 \text{ gausses,} \\ 1 \text{ weber/sq m} &= 10^4 \text{ gausses.} \end{aligned} \tag{37}$$

Thus we establish the relation between the weber and the gauss.

We have now discussed the practical units of electrical and magnetic quantities. At the same time, we have established the laws of force between charges and currents, and the other fundamental relations on which our treatment of the mks units in the text is based. We have not taken up the lucky accident by which the powers of 10 in terms of which the practical units of current and voltage are related to the emu, and the powers of 10 by which the meter and the newton are related to the centimeter and the dyne, are just such that they can be combined into a consistent set of units involving no powers of 10 in the final equations. This is in fact an accident, but when advantage is taken of it, we are led to the set of units taken up in the main discussion in the text. This set of units is still unfamiliar to many readers, who were brought up to use the Gaussian units. The writers believe firmly that the mks units are to be preferred to the Gaussian units, not only for practical calculations, but for all purposes. The equations expressed in terms of mks units are certainly

no more complicated, and in many ways they are simpler, than those expressed in terms of Gaussian units. To that is added the great advantage that, whenever results are to be translated into practice, the units to be used in the mks system are those in practical use, whereas in the Gaussian system they are all unfamiliar, and a conversion table must be used. The only constants that need be remembered in the mks system are ϵ_0 and μ_0, and these have definite physical meanings: ϵ_0 measures the capacity of a condenser whose plates are each 1 m square, and 1 m apart; μ_0 measures the self-inductance, per meter length, of a transmission line consisting of two parallel-plane conductors, each 1 m broad, spaced 1 m apart. (Each of these definitions, of course, implies suitable conductors outside those in question, to prevent fringing of fields.)

APPENDIX III

FOURIER SERIES

Fourier's theorem may be stated as follows: Given an arbitrary function $\varphi(x)$. Then [unless $\varphi(x)$ contains an infinite number of discontinuities in a finite range, or similarly misbehaves itself], we can write

$$\varphi(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \left(A_n \cos \frac{2\pi nx}{X} + B_n \sin \frac{2\pi nx}{X} \right) \tag{1}$$

where

$$A_n = \frac{2}{X} \int_{-X/2}^{X/2} \varphi(x) \cos \frac{2\pi nx}{X} \, dx,$$
$$B_n = \frac{2}{X} \int_{-X/2}^{X/2} \varphi(x) \sin \frac{2\pi nx}{X} \, dx. \tag{2}$$

This equation holds for values of x between $-X/2$ and $X/2$, but not in general outside this range. The series of sines and cosines is called "Fourier's series." There are two sides to the proof of Fourier's theorem. First, we may prove that, if a series of sines and cosines of this sort can represent the function, then it must have the coefficients we have given. This is simple, and we shall carry it through. But second, we could show that the series we so set up actually represents the function. That is, we should investigate the convergence of the series, show that it does converge, and that its sum is the function $\varphi(x)$. This second part we shall omit, merely stating the results of the discussion.

Let us suppose that $\varphi(x)$ is represented by a series as in (1), and ask what values the A's and B's must have if the equation is to be true. Multiply both sides of the equation by $\cos\,(2\pi mx/X)$, where m is an integer, and integrate from $-X/2$ to $X/2$. We then have

$$\int_{-X/2}^{X/2} \varphi(x) \cos \frac{2\pi mx}{X} \, dx = \int_{-X/2}^{X/2} \left(\frac{A_0}{2} \cos \frac{2\pi mx}{X} \right.$$
$$\left. + \sum_n A_n \cos \frac{2\pi nx}{X} \cos \frac{2\pi mx}{X} + \sum_n B_n \sin \frac{2\pi nx}{X} \cos \frac{2\pi mx}{X} \right) dx. \tag{3}$$

But we can easily show by direct integration that

$$\int_{-X/2}^{X/2} \cos \frac{2\pi nx}{X} \cos \frac{2\pi mx}{X} dx = 0$$

if n and m are integers, unless $n = m$, and that

$$\int_{-X/2}^{X/2} \sin \frac{2\pi nx}{X} \cos \frac{2\pi mx}{X} dx = 0$$

if n and m are integers. Thus all terms on the right of (3) are zero except one, for which $n = m$. The first term falls in with this rule, when we remember that $\cos 0 = 1$. This one term then gives us

$$A_n \int_{-X/2}^{X/2} \cos^2 \frac{2\pi nx}{X} dx = A_n \frac{X}{2},$$

as we can readily show. Hence

$$A_n = \frac{2}{X} \int_{-X/2}^{X/2} \varphi(x) \cos \frac{2\pi nx}{X} dx.$$

In a similar way, multiplying by $\sin (2\pi mx/X)$, we can prove the formula for B_n.

We have thus shown that, if a function $\varphi(x)$ is to be represented by a series (1), the coefficients must be given by (2). We shall next make a few remarks about the other part of the problem, the question whether the series so defined really converges to represent the function $\varphi(x)$. In the first place, the series cannot in general represent the function, except in the region between $-X/2$ and $X/2$. For the series is periodic, repeating itself in every period, whereas the function in general is not. Only periodic functions of this period can be represented in all their range by Fourier series. If we try to represent a nonperiodic function, the representation will be correct within the range from $-X/2$ to $X/2$, but the same thing will automatically repeat itself outside the range. Incidentally, we can easily change the range in which the series represents the function. If we merely change the range of integration so as to be from x_0 to $x_0 + X$, where x_0 is arbitrary, the series will represent the function within this range. The case we have used corresponds to $x_0 = -X/2$; another choice frequently made is $x_0 = 0$.

We can also change the value of X, and thereby change the length of the range in which the series is correct. To represent a function through a large range of x, we may use a large value of X. In fact,

as X becomes infinite, the quantities $2\pi nx/X$ for successive n's become arbitrarily close together, and the summations involved in (1) and (2) may be replaced by integrations. Thus, if we let $2\pi nx/X$ equal ω, the interval $d\omega$ between successive values of this quantity will be $d\omega = 2\pi/X$. We may then replace (1) by the integration

$$\varphi(x) = \frac{X}{2\pi} \int_0^\infty (A_n \cos \omega x + B_n \sin \omega x)\, d\omega,$$

where

$$A_n = \frac{2}{X} \int_{-\infty}^{\infty} \varphi(\xi) \cos \omega\xi \, d\xi,$$

$$B_n = \frac{2}{X} \int_{-\infty}^{\infty} \varphi(\xi) \sin \omega\xi \, d\xi.$$

Here we have left out the term in A_0, which becomes negligible in the limit. Combining, we finally have

$$\varphi(x) = \frac{1}{\pi} \int_0^\infty d\omega \int_{-\infty}^{\infty} \varphi(\xi) \cos \omega(\xi - x)\, d\xi. \tag{4}$$

This theorem expresses Fourier's integral theorem, which as we see is merely the limiting form of Fourier's series for infinite value of the period X.

There is an alternative form of Fourier's theorem (1) and (2), expressed in terms of exponential rather than trigonometric functions, which is simpler to write, and for many purposes is more convenient. To derive it, we express the sines and cosines in (1) and (2) in complex exponential form. Grouping together the terms in $e^{j2\pi nx/X}$, and the terms in $e^{-j2\pi nx/X}$, we find without trouble that an equivalent statement of the theorem is

$$\varphi(x) = \sum_{n=-\infty}^{\infty} C_n e^{j2\pi nx/X} \tag{5}$$

where

$$C_n = \frac{1}{X} \int_{-X/2}^{X/2} \varphi(x) e^{-j2\pi nx/X}\, dx. \tag{6}$$

The terms for equal positive and negative values of n combine to give the sines and cosines correctly, and the term for $n = 0$ gives the term in A_0 in (1). Though the C_n's, from their definition, are complex, it is easily shown that, if $\varphi(x)$ is real, C_n and C_{-n} are complex conjugates, and the series itself is real. We can easily write an expres-

sion of Fourier's integral theorem, equivalent to (4), but in exponential language, similar to (5) and (6).

Although the range within which a Fourier series converges to the value of the function it is supposed to represent is limited, as we have seen, to the value X, there is a compensation, in that within this range a Fourier series can be used to represent much worse curves than a power series. Thus a Fourier series can still converge, even though the function has a finite number of finite discontinuities. It can consist, for example, of one function in one part of the region, another in another (in this case, to carry out the integrations, we must break up the integral into separate integrals over these parts, and add them). The less serious the discontinuities, however, the better the convergence. Thus, if the function itself has discontinuities, the coefficients will fall off as $1/n$; whereas, if only the first derivative has discontinuities, the coefficients will fall off as $1/n^2$, etc. Differentiating a function makes the convergence of a series worse, as we can see, for example, if a function is continuous but its first derivative is discontinuous. Then the coefficients fall off as $1/n^2$, but if we differentiate, the coefficients of the resulting series will fall off as $1/n$. There is an interesting point connected with the series for a discontinuous function. If the function jumps from one value u_1 to another u_2 at a given value of x, then the series at this point converges to the mean value, $(u_1 + u_2)/2$.

Problems

1. Expand in Fourier series the function that is equal to $-x$ for x between $-X/2$ and 0, and equal to x for x between 0 and $X/2$.

2. Expand in Fourier series the function that is equal to -1 for x between $-X/2$ and 0, and equal to 1 for x between 0 and $X/2$. See if this series can be found by differentiating the series of Prob. 1 term by term. Consider the convergence of these two series, with reference to their continuity. What happens if we try to differentiate again term by term?

3. Expand in Fourier series the function that is equal to x^2 for x between $-X/2$ and $X/2$. Compute the sum of the first four terms of this series, and see how good an approximation to the function you have.

4. Expand in Fourier series the function that is equal to zero except for x between $-\xi/2$ and $\xi/2$, where $\xi \ll X$, while in this region it equals unity. Discuss the behavior of this function in the limit as ξ becomes zero.

APPENDIX IV

VECTOR OPERATIONS IN CURVILINEAR COORDINATES

Let us assume three orthogonal coordinates q_1, q_2, q_3, so that the three sets of coordinate surfaces, $q_1 =$ constant, $q_2 =$ constant, $q_3 =$ constant, intersect at right angles, though in general the surfaces will be curved. Now let us move a distance ds_1 normal to a surface $q_1 =$ constant. By doing so, q_2 and q_3 do not change, but we reach another surface on which q_1 has increased by dq_1, which in general is different from ds_1. Thus, with polar coordinates, if the displacement is along the radius, so that r is changing, $ds = dr$; but if it is along a tangent to a circle, so that θ is changing, $ds = r\,d\theta$. In general, we have

$$ds_1 = h_1\,dq_1, \qquad ds_2 = h_2\,dq_2, \qquad ds_3 = h_3\,dq_3, \tag{1}$$

where in polar coordinates the h connected with r is unity, but that connected with θ is r. The first step in setting up vector operations in any set of coordinates is to derive these h's, which can be done by elementary geometrical methods. Thus in cylindrical coordinates, where the coordinates are r, θ, z, we have $ds_1 = dr$, $ds_2 = r\,d\theta$, $ds_3 = dz$, so that $h_1 = 1$, $h_2 = r$, $h_3 = 1$. In spherical polar coordinates, r, θ, φ, we have $ds_1 = dr$, $ds_2 = r\,d\theta$, $ds_3 = r \sin\theta\,d\varphi$, so that $h_1 = 1$, $h_2 = r$, $h_3 = r\sin\theta$.

Gradient.—The component of the gradient of a scalar S in any direction is its rate of change in that direction. Thus the component in the direction 1 (normal to the surface $q_1 =$ constant) is

$$\frac{dS}{ds_1} = \left(\frac{1}{h_1}\right)\left(\frac{\partial S}{\partial q_1}\right),$$

with similar formulas for the other components. Thus in cylindrical coordinates we have

$$\mathrm{grad}_r\, S = \frac{\partial S}{\partial r}, \qquad \mathrm{grad}_\theta\, S = \frac{1}{r}\frac{\partial S}{\partial \theta}, \qquad \mathrm{grad}_z\, S = \frac{\partial S}{\partial z}, \tag{2}$$

and in spherical coordinates we have

$$\mathrm{grad}_r\, S = \frac{\partial S}{\partial r}, \qquad \mathrm{grad}_\theta\, S = \frac{1}{r}\frac{\partial S}{\partial \theta}, \qquad \mathrm{grad}_\varphi\, S = \frac{1}{r\sin\theta}\frac{\partial S}{\partial \varphi}. \tag{3}$$

Divergence.—Let us apply the divergence theorem to a small volume element $dV = ds_1\,ds_2\,ds_3$, bounded by coordinate surfaces at q_1, $q_1 + dq_1$, etc. If we have a vector **A**, with components A_1, A_2, A_3 along the three curvilinear axes, the flux into the volume over the face at q_1, whose area is $ds_2\,ds_3$, is $(A_1\,ds_2\,ds_3)_{q_1}$, and the corresponding flux out over the opposite face is $(A_1\,ds_2\,ds_3)_{q_1+dq_1}$, where we note that the area $ds_2\,ds_3$ changes with q_1 as well as the flux density A_1. Thus the flux out over these two faces is

$$\frac{\partial}{\partial q_1}(A_1\,ds_2\,ds_3)\,dq_1 = \frac{\partial}{\partial q_1}(A_1h_2h_3)\,dq_1\,dq_2\,dq_3 = \frac{1}{h_1h_2h_3}\frac{\partial}{\partial q_1}(A_1h_2h_3)\,dv.$$

Proceeding similarly with the other pairs of faces, and setting the whole outward flux equal to div **A** dv, we have

$$\text{div}\,\mathbf{A} = \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial q_1}(A_1h_2h_3) + \frac{\partial}{\partial q_2}(A_2h_3h_1) + \frac{\partial}{\partial q_3}(A_3h_1h_2)\right]. \quad (4)$$

Thus in cylindrical coordinates we have

$$\text{div}\,\mathbf{A} = \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{1}{r}\frac{\partial A_\theta}{\partial \theta} + \frac{\partial A_z}{\partial z} \quad (5)$$

and in spherical coordinates

$$\text{div}\,\mathbf{A} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2A_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\,A_\theta) + \frac{1}{r\sin\theta}\frac{\partial A_\varphi}{\partial\varphi}. \quad (6)$$

Laplacian.—Writing the Laplacian of a scalar S as div grad S, and placing $A_1 = \text{grad}_1\,S$, etc., in the expression for div A, we have

$$\nabla^2 S = \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial q_1}\left(\frac{h_2h_3}{h_1}\frac{\partial S}{\partial q_1}\right) + \frac{\partial}{\partial q_2}\left(\frac{h_3h_1}{h_2}\frac{\partial S}{\partial q_2}\right) + \frac{\partial}{\partial q_3}\left(\frac{h_1h_2}{h_3}\frac{\partial S}{\partial q_3}\right)\right]. \quad (7)$$

Thus in cylindrical coordinates we have

$$\nabla^2 S = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial S}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 S}{\partial\theta^2} + \frac{\partial^2 S}{\partial z^2} \quad (8)$$

and in spherical coordinates

$$\nabla^2 S = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial S}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial S}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 S}{\partial\varphi^2}. \quad (9)$$

Curl.—We apply Stokes's theorem to an approximately rectangular area bounded by q_1, $q_1 + dq_1$, q_2, $q_2 + dq_2$. The line integral of a

vector **A** about the circuit is

$$A_1(q_1,q_2)\,ds_1 + A_2(q_1+dq_1,q_2)\,ds_2 - A_1(q_1,q_2+dq_2)\,ds_1 - A_2(q_1,q_2)\,ds_2.$$

This is approximately equal to

$$\left[\frac{\partial}{\partial q_1}(h_2A_2) - \frac{\partial}{\partial q_2}(h_1A_1)\right] dq_1\,dq_2. \tag{10}$$

Since this must be $\text{curl}_3\,\mathbf{A}\,ds_1\,ds_2$, we have

$$\text{curl}_3\,\mathbf{A} = \frac{1}{h_1h_2}\left[\frac{\partial}{\partial q_1}(h_2A_2) - \frac{\partial}{\partial q_2}(h_1A_1)\right], \tag{11}$$

with similar expressions for the other components. Thus in cylindrical coordinates we have

$$\begin{aligned}
\text{curl}_r\,\mathbf{A} &= \frac{1}{r}\frac{\partial A_z}{\partial \theta} - \frac{\partial A_\theta}{\partial z} \\
\text{curl}_\theta\,\mathbf{A} &= \frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r} \\
\text{curl}_z\,\mathbf{A} &= \frac{1}{r}\frac{\partial}{\partial r}(rA_\theta) - \frac{1}{r}\frac{\partial A_r}{\partial \theta}
\end{aligned} \tag{12}$$

and in spherical coordinates

$$\begin{aligned}
\text{curl}_r\,\mathbf{A} &= \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\,A_\varphi) - \frac{1}{r\sin\theta}\frac{\partial A_\theta}{\partial\varphi} \\
\text{curl}_\theta\,\mathbf{A} &= \frac{1}{r\sin\theta}\frac{\partial A_r}{\partial\varphi} - \frac{1}{r}\frac{\partial}{\partial r}(rA_\varphi) \\
\text{curl}_\varphi\,\mathbf{A} &= \frac{1}{r}\frac{\partial}{\partial r}(rA_\theta) - \frac{1}{r}\frac{\partial A_r}{\partial\theta}.
\end{aligned} \tag{13}$$

APPENDIX V

SPHERICAL HARMONICS

In Chap. III, we considered the solution of Laplace's equation in spherical coordinates, and showed that the solution may be written as a product of a function of r, a function of θ, and a function of φ. The function of φ is $\sin m\varphi$ or $\cos m\varphi$, and the function Θ of θ satisfies the differential equation

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left[l(l+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0. \tag{1}$$

We showed that the function Θ could be written in the form

$$\Theta = \sin^m\theta(A_0 + A_2\cos^2\theta + A_4\cos^4\theta + \cdots)$$

or

$$= \sin^m\theta(A_1 + A_3\cos^3\theta + A_5\cos^5\theta + \cdots), \tag{2}$$

where

$$\frac{A_n}{A_{n+2}} = -\frac{(n+1)(n+2)}{l(l+1) - (m+n)(m+n+1)}. \tag{3}$$

We found further that if l is an integer one or the other of the series in (2) breaks off to form a polynomial. On the other hand, we stated without proof that, if the series does not break off, it leads to a function that becomes infinite when $\cos\theta = \pm 1$, so that such a function cannot be used for expanding solutions of Laplace's equation that remain finite at $\cos\theta = \pm 1$. Thus it is only the polynomial solutions that are usually of use.

These polynomials, expressed in a certain form, with suitably chosen values of the coefficients, are the associated Legendre polynomials. They are ordinarily expressed, not as in (2) in a series of increasing powers of $\cos\theta$, but in a series of descending powers, starting with the term of highest power. In this form, they are defined as

$$\begin{aligned} P_l^m(\cos\theta) = \frac{\sin^m\theta(2l)!}{2^l l!(l-m)!}\Bigg[&(\cos\theta)^{l-m} - \frac{(l-m)(l-m-1)}{2(2l-1)}(\cos\theta)^{l-m-2} \\ &+ \frac{(l-m)(l-m-1)(l-m-2)(l-m-3)}{2\cdot 4(2l-1)(2l-3)}(\cos\theta)^{l-m-4} - \cdots \\ &+ \frac{(-1)^p(l-m)(l-m-1)\cdots(l-m-2p+1)}{2\cdot 4\cdot\cdots(2p)(2l-1)(2l-3)\cdots(2l-2p+1)}(\cos\theta)^{l-m-2p} \\ &+ \cdots\Bigg]. \end{aligned} \tag{4}$$

To show the identity of this expression with (2), we need merely show that the ratio of successive coefficients is as given in (3). To do this, we may identify the term in $(\cos\theta)^{l-m-2p}$ in (4) with the term in $(\cos\theta)^n$ in (2), and the term in $(\cos\theta)^{l-m-2p+2}$ in (4) with that in $(\cos\theta)^{n+2}$ in (2). Doing this, taking the ratio of coefficients from (4), we find easily that it is the same as given by (3), so that (4) forms a legitimate way of writing the solution (2). The constant factor multiplying the series in (4) is chosen to simplify certain formulas.

For certain purposes it is more convenient to let $\cos\theta = x$, and to express the associated Legendre polynomials in terms of x. The equation (1), when expressed in terms of x, is

$$(1-x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + \left[l(l+1) - \frac{m^2}{1-x^2}\right]\Theta = 0. \tag{5}$$

The function (4) can be written in a form that is sometimes useful, and that can be justified directly from the differential equation (5). This is

$$P_l^m(x) = \frac{1}{2^l l!}(1-x^2)^{m/2}\frac{d^{l+m}}{dx^{l+m}}(x^2-1)^l. \tag{6}$$

From the form (6), we can arrive at the expression (4) by expanding $(x^2-1)^l$ in binomial expansion, and carrying out the differentiation term by term. The Legendre functions are the special case of (4) or (6) for which $m = 0$. They are often denoted $P_l(\cos\theta)$ or $P_l(x)$.

The associated Legendre polynomials have many important properties, but for our present purposes the most useful ones are their orthogonality and normalization relations. It can be proved that

$$\begin{aligned}\int_{-1}^{1} P_l^m(x)P_n^m(x)\,dx &= 0 \qquad\qquad \text{if } l \neq n \\ &= \frac{2}{2l+1}\frac{(l+m)!}{(l-m)!}\ \text{if } l = n.\end{aligned} \tag{7}$$

This relation can be used in the following way to satisfy boundary conditions over the surface of a sphere, in solutions of Laplace's equation: As we see from Eq. (3.3), Chap. III, such a solution can be expressed as a function of the angle by

$$\psi = \sum_{lm} P_l^m(\cos\theta)(A_{lm}\sin m\varphi + B_{lm}\cos m\varphi). \tag{8}$$

In this summation, l is to go from 0 to infinity, m from 0 to l (for from our definitions we see that m can never be greater than l). Let us

suppose that the sum (8) is to be a specified function of θ and φ, which we may write $\psi_0(\cos\theta,\varphi)$. Let us then multiply both sides of (8) by a particular function, say $P_p^n(\cos\theta)\sin n\varphi$, and integrate over all solid angles (which amounts to integrating over the surface of the sphere). The element of solid angle is $\sin\theta\,d\theta\,d\varphi$. If we let $x = \cos\theta$, we have $dx = -\sin\theta\,d\theta$. Thus we have

$$\int_{-1}^{1} dx \int_0^{2\pi} d\varphi\, P_p^n(x) \sin n\varphi\, \psi_0(x,\varphi) = \sum_{lm} \int_{-1}^{1} P_l^m(x) P_p^n(x)\, dx$$
$$\int_0^{2\pi} \sin n\varphi (A_{lm} \sin m\varphi + B_{lm} \cos m\varphi)\, d\varphi.$$

Because of the orthogonality of the sines and cosines, the integral over φ on the right is zero unless $n = m$, in which case it is πA_{lm}. Setting $n = m$, the integral over x on the right is zero, because of (7), unless $l = p$, in which case its value is given by (7). Thus we have

$$\frac{2\pi}{2l+1}\frac{(l+m)!}{(l-m)!}A_{lm} = \int_{-1}^{1} dx \int_0^{2\pi} P_p^n(x) \sin n\varphi\, \psi_0(x,\varphi)\, d\varphi. \quad (9)$$

In a similar way we can find a formula for B_{lm}.

APPENDIX VI

MULTIPOLES

The properties of multipoles can be conveniently considered either in rectangular or in spherical coordinates. We shall give both types of discussion, later indicating briefly the relation between them. First we define the dipole and quadrupole moments of a distribution of charges. Suppose we have a collection of point charges, all located near the origin. Let the ith charge be e_i, and let it be located at a point with coordinates ξ_i, η_i, ζ_i. Then the dipole moment is a vector, which we may denote by $\mathbf{p}^{(1)}$, with components

$$p_x^{(1)} = \sum_i e_i\xi_i, \qquad p_y^{(1)} = \sum_i e_i\eta_i, \qquad p_z^{(1)} = \sum_i e_i\zeta_i. \tag{1}$$

It is easy to see that this is consistent with the definition of dipole moment given in the text, by considering for instance a dipole consisting of a charge $+e$ at point d along the x axis, and charge $-e$ at the origin; its dipole moment is then a vector along the x axis, of magnitude ed. Similarly the quadrupole moment is a tensor. (The reader unfamiliar with tensor notation will find a discussion in Appendix V of *Mechanics*, by J. C. Slater and N. H. Frank, McGraw-Hill Book Company, Inc., New York, 1947.) We may denote it by $p^{(2)}$, and its components are

$$\begin{aligned} p_{xx}^{(2)} &= \sum_i e_i\xi_i^2, & p_{yy}^{(2)} &= \sum_i e_i\eta_i^2, & p_{zz}^{(2)} &= \sum_i e_i\zeta_i^2, \\ p_{xy}^{(2)} &= \sum_i e_i\xi_i\eta_i, & p_{yz}^{(2)} &= \sum_i e_i\eta_i\zeta_i, & p_{zx}^{(2)} &= \sum_i e_i\zeta_i\xi_i. \end{aligned} \tag{2}$$

In a similar way one can define an octopole moment $p^{(3)}$, whose components are products of three displacements, and which therefore has three subscripts, etc.

We encounter the various moments of the charge distribution both in finding the field of the distribution at distant points, and in finding its potential energy in an arbitrary potential field. The electrostatic potential ψ at point x, y, z, arising from the charge distribution, by

Coulomb's law, is

$$\psi = \frac{1}{4\pi\epsilon_0} \sum_i \frac{e_i}{\rho_i}, \tag{3}$$

where $\rho_i = \sqrt{(x - \xi_i)^2 + (y - \eta_i)^2 + (z - \zeta_i)^2}$. Assuming that x, y, z are larger than ξ_i, η_i, ζ_i, we may expand $1/\rho_i$, in (3), in Taylor's series in ξ_i, η_i, ζ_i. In doing this, we must take the derivatives of ρ_i with respect to ξ_i, η_i, ζ_i, and then set these quantities equal to zero. We note, however, that the derivative of ρ_i with respect to ξ_i is the negative of its derivative with respect to x, etc. We may then replace the derivatives by the negative derivatives of $r = \sqrt{x^2 + y^2 + z^2}$ with respect to x, y, z. Proceeding in this way, we find that

$$\begin{aligned}\psi = \frac{1}{4\pi\epsilon_0} \Bigg\{ \frac{\sum_i e_i}{r} &- \left[p_x^{(1)} \frac{\partial(1/r)}{\partial x} + p_y^{(1)} \frac{\partial(1/r)}{\partial y} + p_z^{(1)} \frac{\partial(1/r)}{\partial z} \right] \\ &+ \frac{1}{2} p_{xx}^{(2)} \frac{\partial^2(1/r)}{\partial x^2} + \frac{1}{2} p_{yy}^{(2)} \frac{\partial^2(1/r)}{\partial y^2} + \frac{1}{2} p_{zz}^{(2)} \frac{\partial^2(1/r)}{\partial z^2} \\ &+ p_{xy}^{(2)} \frac{\partial^2(1/r)}{\partial x\, \partial y} + p_{yz}^{(2)} \frac{\partial^2(1/r)}{\partial y\, \partial z} + p_{zx}^{(2)} \frac{\partial^2(1/r)}{\partial z\, \partial x} + \cdots \Bigg\}. \end{aligned} \tag{4}$$

Here the first term represents the potential of the total charge, as if it were concentrated at the origin, the next three terms represent the potential of the dipole moment, the next of the quadrupole moment, etc.

The expression (4) is easier to interpret if we note the form of the various derivatives. We have

$$\frac{\partial(1/r)}{\partial x} = -\frac{x}{r^3}, \qquad \frac{\partial^2(1/r)}{\partial x^2} = -\frac{1}{r^3} + \frac{3x^2}{r^5}, \qquad \frac{\partial^2(1/r)}{\partial x\, \partial y} = \frac{3xy}{r^5},$$

with similar formulas for the other derivatives. Thus we see that the potential arising from the dipole moment is

$$\psi = \frac{1}{4\pi\epsilon_0} \frac{[p_x^{(1)}x + p_y^{(1)}y + p_z^{(1)}z]}{r^3}. \tag{5}$$

In the numerator we have the scalar product of the dipole moment and the radius vector; since this is the magnitude of the dipole moment, times r, times the cosine of the angle between, we verify Eq. (5.1) of Chap. III for the potential of a dipole. We note further that the potential arising from a dipole is proportional to $1/r^2$ times a function

of the angle, and by examination of (4) and (5) we find similarly that the potential arising from a quadrupole is proportional to $1/r^3$ times a function of angle, with correspondingly increasing exponents of the inverse powers for the higher multipoles.

The type of expansion we have used can be employed to find the potential energy of a charge distribution in an external field. Let there be a potential function $\psi(x,y,z)$ arising from charges external to the distribution we are considering. Then the potential energy of the distribution in this external field is

$$V = \sum_i e_i\psi(\xi_i,\eta_i,\zeta_i). \tag{6}$$

Expanding $\psi(\xi_i,\eta_i,\zeta_i)$ by Taylor's theorem, this becomes

$$\begin{aligned} V = \left(\sum_i e_i\right)\psi(0) &+ p_x^{(1)}\frac{\partial\psi}{\partial x} + p_y^{(1)}\frac{\partial\psi}{\partial y} + p_z^{(1)}\frac{\partial\psi}{\partial z} \\ &+ \frac{1}{2}p_{xx}^{(2)}\frac{\partial^2\psi}{\partial x^2} + \frac{1}{2}p_{yy}^{(2)}\frac{\partial^2\psi}{\partial y^2} + \frac{1}{2}p_{zz}^{(2)}\frac{\partial^2\psi}{\partial z^2} \\ &+ p_{xy}^{(2)}\frac{\partial^2\psi}{\partial x\,\partial y} + p_{yz}^{(2)}\frac{\partial^2\psi}{\partial y\,\partial z} + p_{zx}^{(2)}\frac{\partial^2\psi}{\partial z\,\partial x} + \cdots. \end{aligned} \tag{7}$$

We note that the dipole terms in the potential energy form the scalar product of the dipole moment and the gradient of the potential, or the negative of the scalar product of dipole moment and electric field. By investigating the change of V, in (7), when the charge distribution is given a translation or a rotation, we can find the forces and torques acting on the charge distribution. We find, of course, that there is no net force acting on a dipole in a uniform field; the force arises only when the field strength varies with position. There is, however, clearly a torque acting on a dipole in a uniform field, tending to force the dipole into parallelism with the field.

We shall now express the formulation of the potential arising from a charge distribution in spherical polar coordinates. Let the coordinates of the charge e_i be r_i, θ_i, φ_i, and the coordinates of the point where we wish to find the potential r, θ, φ. Then, writing x, y, z, ξ_i, η_i, ζ_i, in terms of spherical coordinates, we find that the distance between the point r, θ, φ, and the point r_i, θ_i, φ_i where the ith charge is located is

$$\rho_i = \sqrt{r^2 - 2rr_i[\cos\theta\cos\theta_i + \sin\theta\sin\theta_i\cos(\varphi - \varphi_i)] + r_i^2}. \tag{8}$$

We then substitute in (3) to get the potential. In deriving this,

we shall make use of a very important theorem relating to spherical harmonics, whose proof we shall not give. This is the following:

$$\frac{1}{\rho_i} = \sum_{l=0}^{\infty} \frac{r_i^l}{r^{l+1}} \left[P_l(\cos\theta)P_l(\cos\theta_i) + 2\sum_{m=1}^{l} \frac{(l-m)!}{(l+m)!} P_l^m(\cos\theta)P_l^m(\cos\theta_i)\cos m(\varphi - \varphi_i) \right] \tag{9}$$

if $r > r_i$. This expansion suggests setting up the following quantities, which serve as the various components of multipole moments in spherical coordinates:

$$\begin{aligned} p_0^{(l)} &= \sum_i e_i r_i^l P_l(\cos\theta_i) \\ p_{m,1}^{(l)} &= 2\sum_i \frac{(l-m)!}{(l+m)!} e_i r_i^l P_l^m(\cos\theta_i)\cos m\varphi_i \\ p_{m,2}^{(l)} &= 2\sum_i \frac{(l-m)!}{(l+m)!} e_i r_i^l P_l^m(\cos\theta_i)\sin m\varphi_i. \end{aligned} \tag{10}$$

We shall point out the nature of these quantities later, and shall show that those for $l = 1$ describe the dipole moment, those for $l = 2$ the quadrupole moment, etc. In terms of them, the potential is then

$$\psi = \frac{1}{4\pi\epsilon_0}\sum_l \frac{1}{r^{l+1}} \left\{ p_0^{(l)}P_l(\cos\theta) + \sum_{m=1}^{l} [p_{m,1}^{(l)}P_l^m(\cos\theta)\cos m\varphi + p_{m,2}^{(l)}P_l^m(\cos\theta)\sin m\varphi] \right\}. \tag{11}$$

Let us now examine the expression (11), to find its significance. The term in $l = 0$ is very simple. Remembering, as follows from Appendix V, that $P_0(\cos\theta) = 1$, it is

$$\frac{1}{4\pi\epsilon_0}\frac{\sum_i e_i}{r}$$

expressing the potential of the total charge, as in the first term of (4). For the term in $l = 1$, we have $P_1(\cos\theta) = \cos\theta$, $P_1^1(\cos\theta) = \sin\theta$.

Then

$$
\begin{aligned}
p_0^{(1)} &= \sum_i e_i r_i \cos \theta_i = \sum_i e_i \zeta_i \\
p_{1,1}^{(1)} &= \sum_i e_i r_i \sin \theta_i \cos \varphi_i = \sum_i e_i \xi_i \\
p_{1,2}^{(1)} &= \sum_i e_i r_i \sin \theta_i \sin \varphi_i = \sum_i e_i \eta_i .
\end{aligned} \tag{12}
$$

Thus these three quantities are the components of the dipole moment, as we have defined them previously. Substituting in (11), we find that the term for $l = 1$ reduces to exactly the expression (5), when we remember that $x = r \sin \theta \cos \varphi$, $y = r \sin \theta \sin \varphi$, $z = r \cos \theta$. In a similar way we can identify the terms in $l = 2$ in (11) with the quadrupole terms in the rectangular case, though this identification is not so simple as with the dipole, and involves a paradox, in that there are only five components of the form given in (10), whereas we found six components of the quadrupole moment in (2). The paradox becomes resolved when we write out the potential in detail in both coordinate systems. Proceeding in a similar way, we see that the term of (11) corresponding to each value of l expresses the potential arising from a particular multipole.

APPENDIX VII

BESSEL'S FUNCTIONS

Bessel's equation is

$$\frac{1}{z}\frac{d}{dz}\left(z\frac{dZ}{dz}\right) + \left(1 - \frac{m^2}{z^2}\right)Z = 0, \tag{1}$$

and its solutions are $Z = J_m(z)$, $Z = N_m(z)$. By expanding Z in power series, we may show that

$$J_m(z) = \frac{1}{m!}\left(\frac{z}{2}\right)^m - \frac{1}{(m+1)!}\left(\frac{z}{2}\right)^{m+2} + \frac{1}{2!(m+2)!}\left(\frac{z}{2}\right)^{m+4} \cdots \tag{2}$$

where the coefficient of the first term is chosen according to an arbitrary convention. There is no similarly simple expansion for the Neumann function $N_m(z)$; it requires not only terms in positive and negative powers of z, but also logarithmic terms. For $N_0(z)$, the leading term for small values of z is the logarithmic term:

$$\lim_{z\to 0} N_0(z) = \frac{2}{\pi}(\ln z - 0.11593).$$

For the higher values of m, the term in inverse powers of z is the leading term for small z:

$$\lim_{z\to 0} N_m(z) = -\frac{(m-1)!}{\pi}\left(\frac{2}{z}\right)^m, \qquad m > 0. \tag{3}$$

For values of z large compared with m, the Bessel and Neumann functions have asymptotic expressions as follows:

$$\begin{aligned}\lim_{z\to\infty} J_m(z) &= \sqrt{\frac{2}{\pi z}}\cos\left(z - \frac{2m+1}{4}\pi\right),\\ \lim_{z\to\infty} N_m(z) &= \sqrt{\frac{2}{\pi z}}\sin\left(z - \frac{2m+1}{4}\pi\right).\end{aligned} \tag{4}$$

An important relation connecting the Bessel and Neumann functions is

$$N_{m-1}(z)J_m(z) - N_m(z)J_{m-1}(z) = \frac{2}{\pi z}.$$

An important relation gives $J_m(z)$ as an integral:

$$J_m(z) = \frac{1}{2\pi j^m} \int_0^{2\pi} e^{jz\cos w} \cos (mw)\, dw,$$

where $j = \sqrt{-1}$.

In addition to these relations, there are a group of important relations holding equally well for either the Bessel or Neumann functions, and which we shall therefore state in terms of $Z_m(z)$, meaning by this either $J_m(z)$ or $N_m(z)$. For the Bessel functions, most of these relations can be proved without difficulty from the series representation. The relations are more general, however, in that they apply to the Neumann functions as well, and can be proved directly from the properties of the differential equation. They are the following:

$$\begin{aligned}
Z_{m-1}(z) + Z_{m+1}(z) &= \frac{2m}{z} Z_m(z) \\
Z_{-m}(z) &= (-1)^m Z_m(z), \qquad \text{if } m \text{ is integral} \\
\frac{d}{dz} Z_m(z) &= \frac{1}{2}[Z_{m-1}(z) - Z_{m+1}(z)] \\
\frac{d}{dz}[z^m Z_m(z)] &= z^m Z_{m-1}(z) \\
\frac{d}{dz}[z^{-m} Z_m(z)] &= -z^{-m} Z_{m+1}(z) \\
\int Z_1(z)\, dz &= -Z_0(z) \\
\int zZ_0(z)\, dz &= zZ_1(z) \\
\int Z_0^2(z) z\, dz &= \frac{z^2}{2}[Z_0^2(z) + Z_1^2(z)] \\
\int Z_m^2(z) z\, dz &= \frac{z^2}{2}[Z_m^2(z) - Z_{m-1}(z) Z_{m+1}(z)].
\end{aligned}$$

In Chap. XII, Sec. 2, we introduced spherical Bessel and Neumann functions, defined by the equations (2.1) of that chapter. We showed in Eq. (2.2) of that chapter that they can be expressed in terms of algebraic and trigonometric functions; we found their asymptotic behavior in (2.3), and their behavior for small z in (2.4), all of that chapter. These properties can be easily derived from the relations we have already given. For the series representation, we must modify (2) for nonintegral values of m, replacing $m!$ where it appears in that equation by the gamma function $\Gamma(m + 1)$, where the gamma

function satisfies the functional relation

$$\Gamma(m+1) = m\Gamma(m),$$

and where $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. With this modification, (2) and (3) lead to Eq. (2.4), Chap. XII. The asymptotic behavior for large z given in Eq. (2.3), Chap. XII, follows directly from (4). To prove the trigonometric nature of the functions, we can first find $j_0(z)$ and $n_0(z)$ directly, showing that as given in (2.2), Chap. XII, they are given by $(\sin x)/x$ and $(\cos x)/x$, respectively. To do this, we can substitute the corresponding functions $J_{1/2}(z)$ and $N_{1/2}(z)$, defined from them by (2.1), Chap. XII, directly in Bessel's equation (1), and show that it is satisfied. Then we can derive the functions of higher index from these by the relation

$$\frac{d}{dz}[z^{-m}Z_m(z)] = -z^{-m}Z_{m+1}(z),$$

which we have given above, leading immediately to the algebraic and trigonometric nature of the higher functions, as well as to the explicit values for the functions.

SUGGESTED REFERENCES

Electromagnetic theory is too large a subject to be treated completely in a single volume, and we give a few references in the present section, which the student familiar with this book can refer to, without too great difficulty. In the first place, the reader of inadequate preparation may wish to review his elementary electromagnetism and optics, and can use, for example, *Introduction to Electricity and Optics*, by N.H. Frank (McGraw-Hill).

For general mathematical training and background, the student will first want texts on advanced calculus, such as *Treatise on Advanced Calculus*, by P. Franklin (Wiley), or *Advanced Calculus*, by E.B. Wilson (Ginn). More advanced texts on analysis will be helpful, such as *Mathematical Analysis*, by Goursat and Hedrick (Ginn), *Partielle Differentialgleichungen der Physik*, by Riemann and Weber (Rosenberg), or *Higher Mathematics*, by R.S. Burington and C.C. Torrance (McGraw-Hill). Several texts on mathematics for special purposes are valuable: *Applied Mathematics for Engineers and Physicists*, by L.A. Pipes (McGraw-Hill); *The Mathematics of Physics and Chemistry*, by H. Margenau and G.M. Murphy (Van Nostrand); and *Mathematical Methods for Engineering*, by T. von Kármán and M.A. Biot (McGraw-Hill).

Among more specialized mathematical texts in the fields of particular importance to electromagnetism are *Introduction to Higher Algebra*, by M. Bôcher (Macmillan); *Ordinary Differential Equations*, by E.L. Ince (Dover); *Fourier Series and Spherical Harmonics*, by W.E. Byerly (Ginn); *Newtonian Potential Function*, by B.O. Peirce (Ginn); *Fourier Series and Boundary Value Problems*, by R.V. Churchill (McGraw-Hill); *Vector Analysis*, by H.B. Phillips (Wiley); and *Vector and Tensor Analysis*, by H.V. Craig (McGraw-Hill). The standard volumes of tables, *A Short Table of Integrals*, by B.O. Peirce (Ginn) and *Funktionentafeln*, by Jahnke and Emde (Dover), will be found invaluable for detailed assistance in calculation. For definite integrals that are not given in these books, *Nouvelles Tables d'Intégrales Définies*, by Bierens de Haan (Stechert), will be found a source of much information.

A number of general texts on theoretical physics cover electromagnetic theory among other topics. Among these we may mention

Introduction to Theoretical Physics, by L. Page (Van Nostrand); *Theoretical Physics,* by G. Joos (Stechert); *Introduction to Theoretical Physics,* by A. Haas (Constable); *Introduction to Mathematical Physics,* by R.A. Houstoun (Longmans); and *Principles of Mathematical Physics,* by W.V. Houston (McGraw-Hill). Two longer treatises on theoretical physics, in several volumes, may also be mentioned: *Introduction to Theoretical Physics,* by M. Planck (Macmillan), an English translation of a well-known German text, and *Einführung in die theoretische Physik,* by C. Schaefer (De Gruyter). The last two works go a good deal more into detail than is possible in the present book.

Next we come to a number of references dealing with the various branches of electromagnetic theory. Among general references, two standard works are *Classical Electricity and Magnetism,* by Abraham and Becker (Blackie), and *The Mathematical Theory of Electricity and Magnetism,* by J.H. Jeans (Cambridge). Treatments on approximately the scale of the present text are given in *Principles of Electricity,* by L. Page and N.I. Adams (Van Nostrand), *Electric Oscillations and Electric Waves,* by G.W. Pierce (McGraw-Hill), *Principles of Electricity and Electromagnetism,* by G.P. Harnwell (McGraw-Hill), and *Static and Dynamic Electricity,* by W.R. Smythe (McGraw-Hill). More advanced points of view are presented in *Electromagnetic Theory,* by J.A. Stratton (McGraw-Hill), and *Electromagnetic Waves,* by S.A. Schelkunoff (Van Nostrand)

A number of texts handle more specialized problems. In electron theory, and the theory of dielectrics and magnetic materials, there are *Theory of Electrons,* by H.A. Lorentz (B.G. Teubner); *Theory of Electric and Magnetic Susceptibilities,* by J.H. Van Vleck (Oxford); *Polar Molecules,* by P. Debye (Chemical Catalog Co.); and *Introduction to Ferromagnetism,* by F. Bitter (McGraw-Hill). The theory of wave guides and microwaves is handled in a number of recent texts, including *Microwave Transmission,* by J.C. Slater (McGraw-Hill); *Fields and Waves in Modern Radio,* by S. Ramo and J.R. Whinnery (Wiley); and *Hyper and Ultrahigh Frequency Engineering,* by R.I. Sarbacher and W.A. Edson (Wiley). The subject of optics is handled from the standpoint of electromagnetic theory in *Lehrbuch der Optik,* by K. Försterling (S. Hirzel), and *Optik,* by M. Born (Springer). Other standard treatments of optics are given in *Fundamentals of Physical Optics,* by F.A. Jenkins and H.E. White (McGraw-Hill), and in *Physical Optics,* by R.W. Wood (Macmillan).

INDEX

L

M

N

O

P

Q

R

S

T

U

V

W

X